Paper prepared for the 22nd Biennial Conference of the Society for Multivariate Analysis in the Behavioural Sciences at the London School of Economics (17-19 July 2000)

Conceptual issues arising from a comparability study relating IGCSE grading standards with those of GCSE via a reference test using a multilevel model

Trevor Dexter and Alf Massey
Research and Evaluation Division
University of Cambridge Local Examinations Syndicate
1 Hills Road
Cambridge
CB1 2EU

Introduction

This paper was produced for the 22nd Biennial Conference of the Society for Multivariate Analysis in the Behavioural Sciences at the London School of Economics (17-19 July 2000). The presentation at the conference looked at conceptual issues arising from a comparability study relating IGCSE grading standards with those of GCSE via a reference test using a multilevel model.

The format of this paper consists of a summary of what was said in the presentation followed by a version of the research upon which the presentation was based.

Summary of presentation

Grading standards in the June 1997 IGCSE (International General Certificate of Secondary Education) examinations in fourteen subjects were contrasted with those in similar British GCSE examinations. The IGCSE, provided by the University of Cambridge Local Examinations Syndicate (UCLES), is a curriculum for the 14-16 group used in international schools throughout the world with examinations leading to an internationally recognised certificate equivalent in standard to GCSE. Comparisons used a 'Calibration Test' (containing items testing verbal, numerical and spatial reasoning) to control for variations in general ability between candidates for different examinations. It was administered to 1664 IGCSE candidates from 39 international schools, and to 3656 UK GCSE candidates from 23 comprehensive and 20 independent/selective schools. As candidates are nested within schools, multilevel models were used to analyses the data. The models had exam grade (A*=8,..,U=0) as the response with terms for standardised Calibration Test score, GCSE comprehensive students, GCSE independent/selective students and interactions (IGCSE is the base). Terms for gender and the non-English home language of IGCSE candidates were included along with interactions. There was no suggestion that IGCSE grading standards in any of the subjects investigated were more or less lenient than those established in the GCSE examinations. This paper concerns some of the conceptual issues that arose whilst interpreting the results of the multilevel modelling.

On average IGCSE candidates achieved slightly higher grades than GCSE candidates of equivalent ability from comprehensive schools, but lower grades than GCSE candidates in independent/selective schools, with their regression lines converging for the brightest students (a similar pattern held across subjects). The IGCSE regression line lies between that of the GCSE comprehensive schools and the GCSE independent/selective schools. The GCSE school type effect is interesting as IGCSE can not simply be expected to be concordant to a single GCSE line. Gender differences existed with IGCSE girls doing better than boys in subjects like English, English Literature, French (English home language IGCSE candidates only), Biology and Chemistry. Similar sex differences were evident in both the GCSE comprehensive and independent/selective groups but the UK comprehensive schools girls also showed superiority in the humanities and art. IGCSE home language English students only performed better on English examinations with IGCSE non-English home language students performing better in Mathematics, Physics, Computer Studies and French, reflecting linguistic influences on performance in both the Calibration Test and the IGCSE examinations.

Most 'reference test' comparability studies effectively test whether regression lines are coincident, maybe using additional controlling variables. Here there was no simple hypothesis of the regression lines being co-incident. There were variations in the relationship between the achieved examination grades and Calibration Test scores for GCSE school type, gender across exams and IGCSE candidates' language background. The implication of such variations is that it moves this comparability study away from testing a single null hypothesis, towards a description of the relationship between variables in different groups from which judgements concerning standards are made.

The research upon which the presentation was based

by A.J.Massey and T.Dexter (1998)
Research & Evaluation Division, University of Cambridge Local Examinations Syndicate, 1 Hills Road, Cambridge, CB1 2EU

Grading Standards and the IGCSE examination

The International General Certificate of Secondary Education (IGCSE) examination provided by the University of Cambridge Local Examination Syndicate (UCLES) is designed as a two year curriculum programme for the 14-16 age group, tailored to international needs. It has been especially welcomed in international schools throughout the world, where it helps to provide a common curriculum, facilitating transfer of geographically mobile students and teachers (UCLES, 1995). End of course examinations lead to a certificate which is internationally recognised as equivalent in standard to British General Certificate of Secondary Education (GCSE) examinations, which UCLES also provides via the Midland Examining Group (MEG)[1]. As in the GCSE, IGCSE is in essence[2] a 'single subject' examination, in that candidates may choose to enter for one or more of the subject examinations serving a wide selection of syllabuses. As in GCSE, Grades are awarded on a nine point scale (Grades A*, A through G and Unclassified). Whilst GCSE examinations are taken in June each year, IGCSE examinations are set in June and November, with schools in the Northern hemisphere largely utilising the June examination and those from the Southern hemisphere taking the examinations in November; in both cases towards the end of the academic year.

The IGCSE syllabuses and examinations were designed to utilise the experience of GCSE and to provide a balanced curricular experience with an international flavour. They provide for positive achievement across the ability range at age 16 by means of a choice between a 'Core' and an 'Extended' curriculum in most subjects, with differentiated examination papers targeted at these. Core examinations are targeted at pupils expected to obtain grades D-G and are limited to grade C, whilst Extended papers are targeted at pupils likely to reach grades A-C. A range of assessment methods are employed, appropriate to the knowledge and skills assessed in each subject. Examination papers are in the main set and marked in Cambridge by experienced examiners, although a school based coursework component is available as (an optional) part of the assessment in subjects where it can enhance validity by extending the range of assessments possible; provided teachers who have received special training from UCLES are available to manage and assess it. But whilst IGCSE syllabuses have built on the experience of GCSE syllabuses designers in the same subjects, share many of their features and are judged to make equivalent intellectual demands on students they, and the examinations which serve them, are clearly similar, rather than identical, to the GCSE.

As in GCSE, standard setting in IGCSE examinations has been based on the judgements of experienced examiners, who are also familiar with the quality of work required of British students. The examiners have been charged with the responsibility of ensuring that equivalent standards to those of GCSE are set and maintained, year on year, in IGCSE.

[1] Following reorganisation and mergers, from 1999 onwards UCLES' participation in UK examinations will be via the Oxford, Cambridge and Royal Society of Arts (OCR) examining body.
[2] Candidates offering a proscribed selection of seven or more syllabuses are also eligible for a 'grouped certificate', the International Certificate of Education (ICE).

The IGCSE examinations began in 1988 and now attract substantial entries. It is clearly the responsibility of the examining body to evaluate the equivalence of standards asserted, to reassure those who take IGCSE examinations or take its standards on trust when selecting for employment or further education. But professional judgements apart, it is not easy to provide evidence bearing on equivalence of examination standards. Comparisons of the distributions of grades awarded are clearly nonsensical: the IGCSE entry, largely drawn from pupils in fee paying schools distributed throughout the world which cater for a mixture of expatriate (not by any means all of British origin) and local students, is not necessarily in any sense of equivalent 'ability' to the entry for any given MEG GCSE syllabus, drawn from a mixture of state and fee paying UK schools. Given that the syllabuses in a given subject are not the same, students must be prepared by their teachers for one examination or the other, and it is thus not feasible to ask one group to take the other's assessments, even experimentally, without a bias arising. This rules out direct comparisons of marks or grades for the same students. We are thus forced towards an indirect comparison and should recognise the methodological and conceptual implications of this.

The need for calibration and ways of achieving it

A 'common test' taken by all, can serve as a 'common yardstick' against which to calibrate groups of candidates taking different achievement tests or examinations, so that the 'equivalence' of the marks or grades awarded in such examinations may be investigated.

The use of such 'reference tests' tests to monitor comparability of standards is long established (Wrigley, Sparrow & Inglis, 1967; Willmott, 1997) although the methodology has been criticised (Murphy, Wilmut and Wood, 1996). There has been substantial debate about the relative merits of using subject-based tests rather than general tests as the basis for comparisons, with subject-based tests often said to be more relevant to (highly correlated with) academic achievement. But subject-based tests are also inherently more likely to be differentially biased (Newbould and Massey, 1979) against one or more of the examinations being compared, with pupils from one curricular route disadvantaged relative to another on some questions because their learning programmes have emphasised different features of the subject. This would mean that we should not expect equivalent pupils to obtain the same scores on a common test, rendering the null hypothesis untenable and making it difficult or impossible to interpret reference test data. Developing a subject-based reference test for comparing standards between two given examinations and establishing its suitability is in itself a substantial research undertaking, where there can be no guarantee of success.

General tests inform a general question: how do those with given levels of general ability (as measured by the reference test used) perform on the various other tests or examinations being compared. The operational definition of equivalent standards is therefore that students with a given general ability 'score', who are following 'comparable' courses leading to different achievement tests or examinations, might be expected to exhibit similar achievement and to obtain equivalent marks or grades. The main assumptions are thus that the reference test is an equally relevant and unbiased measure with respect to the achievement domains compared and that there is no reason to believe that one of the curricular regimes which underpin the examinations is more effective in converting aptitude into achievement than another. Data analysis is conceptually quite simple, even though it may in practice prove rather sophisticated because of the need to control for factors which may contravene the assumptions above. Linear regression models are used to test the null hypothesis that pupils with similar calibration test scores should, on average, achieve

ostensibly equivalent achievement test marks/ grades/ levels etc. from their (different) test regimes.

A general ability tests thus has some advantages as a reference test. Distanced as its content is from the curriculum based tests to be compared, it is less likely to be seen as biased by giving any advantage to pupils prepared under one regime or another, so that the null hypothesis above remains tenable. Relevance, or the degree of association between the reference test and the achievement measures, is another matter. It is indeed helpful if a reference test correlates highly with the achievement measures, because this adds power to statistical comparisons controlling for variations in the abilities it measures, as well as bolstering face validity for its use in monitoring standards. But it is more important that the reference test relates to the different achievement measures in a similar way than that the correlation is high. Where it correlates better with one achievement measure than another the statistical basis for comparisons will be severely eroded and such variations are at least as, and perhaps more, likely in subject-based reference tests. Resourcing is a further major issue. The costs of developing a suitable subject-based reference test for a given comparison are likely to be substantial and the development of a series of such measures for the full spectrum of subject examinations, as would be required to monitor IGCSE's compliance with GCSE standards, is prohibitive.

Whilst not wishing to claim too much for the use of general ability measures in monitoring standards, this approach can serve as a prima facie basis for comparison in many circumstances. In the context of the scrutiny of public examination standards, general ability reference tests may perhaps best be seen as a cost-effective (Nuttall, 1971) screening device: best used to discover where further scrutiny of standards may be worthwhile and to support professional judgements about the desirability and extent of remedial action (Willmott, 1980). This seems to fit our need to evaluate the equivalence of IGCSE and GCSE grading standards across the full range of subjects.

The 'Calibration Test'

In order to provide a convenient measure of general ability for a wide-ranging programme of research comparing standards in a variety of achievement tests, including those in the several suites of examinations provided under the auspices of UCLES, the Research and Evaluation Division at UCLES have developed two parallel forms (A and B) of a short (30 minute/ 60 item), objective 'Calibration Test'. This is designed to be easily administered (by pupils' own teachers), machine marked and to correlate with achievement measures in a wide range of subjects. Development sought a difficulty level making the test effective in the full ability range at about age 16 in English speaking populations, whilst probably also being suitable for use with older/younger and/or more heavily selected groups: such as secondary school pupils aged about 14+, or the relatively able UK students opting for courses leading to General Certificate in Education (GCE) A Level examinations at age 18. The test was intended to yield a total 'general ability' score.

The Calibration Test is not intended to break new psychometric ground. For the most part it uses item types often found in similar tests; requiring verbal, numerical and spatial reasoning. It is also speeded: students are asked to work quickly and it is not anticipated that all will find it possible to finish within the allotted time. The intention is to provide a brief, cost-effective and reliable group test of general ability, readily available for large-scale use in research monitoring standards and investigating equivalence. Such a test might assist with monitoring standards between equivalent tests or examinations in the same subject testing

different syllabuses (perhaps set by different boards or in different countries); or with the more complex problems of equivalence between achievement tests set in different years; or different subjects; or the alignment of standards between tests normally set to groups of different ages or abilities.

The data collected for this study (described below) provide the first large scale application of the Calibration Test and the test's design, development and evidence relating to reliability and predictive and construct validity are discussed in detail by Massey et al (1998). In summary, it proved to be well targeted and to differentiate effectively (the mean and standard deviation of GCSE candidates' scores being 37.7 (63%) and 9.8 (16%) respectively) and reliable (Coefficient Alpha for Form A being 0.9). Correlations observed between this general ability test and examination grades[3] were also high, often rivalling those found for subject based tests of this length. For instance the correlations with the six most popular GCSE syllabuses 'sampled' were 0.74 (a Maths syllabus with coursework), 0.68 (Geography), 0.66 (English), 0.63 (Science), 0.61 (French) and 0.56 (History). These were not untypical. Higher correlations were observed with other syllabuses and a reasonable correlation (0.53) was even recorded with GCSE grades for Art and Design! In short the Calibration Test appears to be fit for the purpose for which it was intended. However this does not preclude the need to establish that the relationship between test scores and examination results are equally relevant to the specific examinations compared, or to consider how the relationship may vary in different sub-groups of candidates (for instance boys and girls) and how this might affect expected examination outcomes. These matters will be considered further below.

Data collection

The comparisons we wish to make require the collection of data from sufficient numbers of candidates for each of a selection of IGCSE and GCSE syllabuses spanning the range of subjects examined. Comparisons can then be made between grading standards in examinations in the same subjects drawn from the two suites. These comparisons require groups of candidates from each subject examination involved which span the full ability range, so that the regression of examination grade on calibration test score can be soundly estimated in each case. But 'representative' samples are not required, as there is no need to estimate the means or variance of the grades awarded to the 'populations' entering the different examinations. Neither is there any need for the calibre of the group taking one examination to be equivalent to the calibre of the candidates for the examination it is compared with, as the Calibration Test data is gathered for the express purpose of 'controlling for' such disparities.

[3] Because the substantive interest lay in comparing syllabuses and because examinations often include optional papers which are judgementally equated in the grading process, it is necessary to use the grades, rather than underlying marks, as the basis for the analysis of the relationships between the Calibration Test and achievement in GCSE examinations. For this purpose grades have been converted to a 0-8 scale. Bardell, Forrest and Shoesmith (1978) point out that this widespread practice is crude and assumes that grades are equally spaced but suggest that experience indicates that final outcomes are little affected by using straightforward numerical conversion without normalisation.

MEG GCSE data

The GCSE data collected came from a range of MEG GCSE syllabuses examined in the Summer of 1997. GCSE examinations mark the end of the compulsory schooling in the United Kingdom[4], where almost all pupils (apart from a small proportion with Special Educational Needs) sit GCSE examinations in a range of academic subjects (varying according to choices by schools and individual pupils) towards the end of Year 11; when most are about 16 years of age. As well as the choices available between optional subjects in the school's own curriculum and between alternative syllabuses offered by MEG in most subjects, schools may also choose between different ranges of syllabuses offered by several examining bodies besides MEG, so the combinations of different subject examinations sat by pupils in UK Schools are legion and the characteristics of pupils taking one syllabus can vary greatly from those in another.

This study sought to maximise the efficiency of data collection by contacting the 163 UK schools which offered three[5] or more of six relatively popular syllabuses (English 1510; French 1525; Geography (Bristol Project) 1588; History (1914-Present) 1607; Maths (With Coursework) 1661; Science (Salters) Double Award 1774) and entered at least twenty candidates for one of them. In addition 28 further schools were approached which entered candidates for all three 'separate science' syllabuses (Biology 1780, Chemistry 1781 and Physics 1782) and had at least 50 candidates in one of them; these syllabuses being of special interest. Schools were asked to administer the Calibration Test to all Year 11 pupils entering GCSE examinations within the two months immediately before these began. Of the 191 schools approached 47 agreed to participate and were supplied with test materials. In the event 43 (22.5%) completed the testing and supplied data. Given the onerous requirement, involving many or all teachers in a school, at a critical and busy phase in the school year, this response rate is perhaps as high as might reasonably be expected. Pupils were asked to indicate their gender and whether they normally spoke English at home[6] as well as completing the Calibration Test. The data collected were subsequently matched (via the examination database) with grades awarded in a wide range of MEG syllabuses, including those mentioned above, for a total of 3,656 pupils who had attempted Form A of the Calibration Test[7].

Thus whilst (for lack of a better term) below we will refer to the schools and candidates providing GCSE data as a 'sample', this is really a misnomer. The group contains pupils spanning almost the full range of ability, drawn from 43 UK secondary schools, but it is not representative of any particular population (and did not need to be for the purpose of this research). Comparison of the distributions of GCSE grades achieved by pupils in the 'sample' for the 'target syllabuses' listed above with those for the full MEG entry, suggests that the sample includes a higher proportion of abler students than is typical. The explanation may lie partly in the inclusion of schools entering candidates for the separate science syllabuses, which are normally considered suitable for abler candidates. The MEG grade distributions for the separate sciences confirm the highly selective nature of their entries compared to most subjects and the candidates for separate sciences in our sample are exceptional even for these syllabuses.

---

[4] Except Scotland.

[5] Excluding those offering only 1510, 1525 and 1607.

[6] 30 candidates did not provide this information.

[7] Ten of the schools were asked to administer Forms A and B of the Calibration Test to alternate pupils to provide equating data. Pupils from these schools attempting Form A are included in this sample, whilst a further 373 pupils attempted Form B.

IGCSE data

IGCSE data collection targeted all schools with a June 1997 subject entry (the cumulative total of candidate entries for all subjects) of at least 180 from the 30 countries world-wide with the largest IGCSE entries. In all 90 IGCSE schools met this criterion and were invited to take part by testing all their IGCSE candidates shortly before their examinations. The 45 which agreed to do so were supplied with test materials and 39 (43%) returned completed Calibration Tests for a total of 1,664 pupils whose data could, subsequently, be matched with their IGCSE results. Again candidates were asked to supply information concerning gender and the language[8] spoken at home. Their geographical origins are shown in table 1. Again it should be recognised that whilst these candidates are not a representative sample this does not invalidate the analyses required for this study.

Table 1 Number of IGCSE schools from each country

Country	*no of schools*	*no of candidates*
United Arab Emirates	7	371
Brazil	1	14
Canada	1	7
Columbia	1	62
Cyprus	1	23
Germany	1	43
Greece	1	38
Hong Kong	1	97
India	2	67
Italy	1	27
Kenya	7	266
Kuwait	1	50
Malawi	1	38
Netherlands	1	44
Philippines	1	4
Singapore	2	195
Spain	3	96
Switzerland	3	119
Thailand	1	26
Zambia	2	77
Total	39	1664

Some Features of Calibration Test Scores

Means and standard deviations of total scores on Form A of the Calibration Test for both IGCSE and GCSE pupils are shown in table 2, which further details the performance of males and females, those who indicated that they did (EHL) or did not (Non-EHL) speak English at home and (for GCSE only) those attending state maintained comprehensive and selective schools and independent schools. The overall distributions of total scores for the IGCSE and GCSE candidates included in the samples are shown in figure 1, which reveals a slight positive skew in both cases, in keeping with these relatively high calibre samples.

Overall, GCSE the 3,656 GCSE candidates in this study obtained a Calibration Test mean score (37.7) close to that of the 1,664 IGCSE candidates (36.7), with a slightly wider spread of marks and a slightly less positively skewed distribution than the IGCSE group.

[8] 79 IGCSE candidates declined to provide this information.

Figure 1 IGCSE and GCSE Sample Calibration Test Total Score Distributions

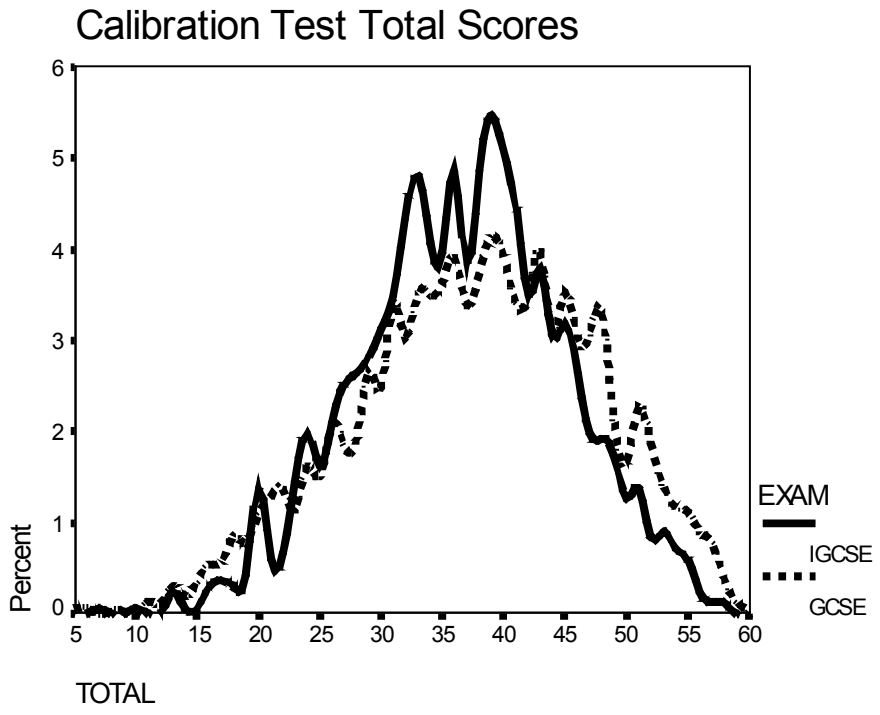


Table 2 Mean and sd of Calibration Test Form A Scores

		n	Total mean	Score sd
GCSE Comprehensive	m	1.088	35.4	9.0
	f	1,219	32.3	8.8
	all	2,307	33.8	9.0
GCSE Independent	m	497	45.5	7.0
	f	474	42.4	7.1
	all	971	44.0	7.2
GCSE Selective	m	149	45.2	6.4
	f	229	46.0	6.7
	all	378	45.7	6.6
GCSE: All Schools	m	1,734	39.1	9.6
	f	1,922	36.4	9.9
	EHL	3,523	37.9	9.7
	Non-EHL	101	34.3	11.3
	all	3,656	37.7	9.8
IGCSE	m	885	37.6	8.3
	f	779	35.6	8.2
	EHL	946	36.7	8.5
	Non-EHL	639	36.7	7.8
	all	1664	36.7	8.3

Gender effects

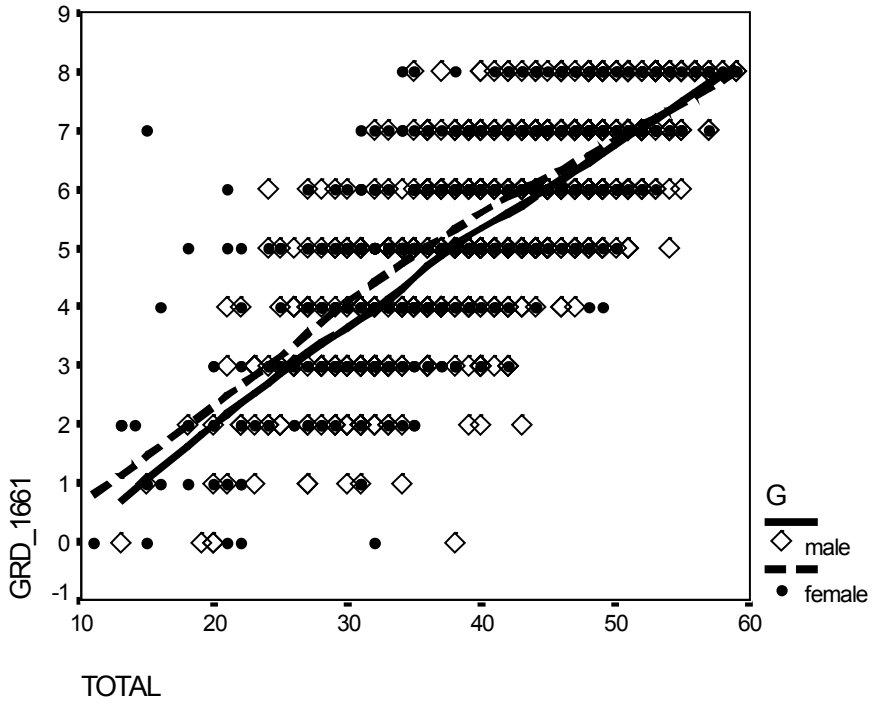
Gender differences in Test scores are evident in both examination groups; unsurprisingly, as such differences are common in both ability and achievement tests (Gipps and Murphy, 1994, provide a thorough review of the complex literature in this area). The data for comprehensive pupils is perhaps the best guide on this issue, as it is quite possible that selection effects may affect the calibre of male and female pupils attending particular international schools or independent or selective schools; factors such as fee levels, parents'

attitudes to boarding for boys and girls, and the proximity, range and selection policies of local competitors may influence their recruitment. The variations in mean scores observed here for male and female students in such schools may well reflect selection effects and should thus not be taken as 'representative'. But there seems no reason why the boys and girls attending comprehensives should be untypically balanced in ability because of selection effects, even though the sample included a relatively high proportion of girls due to the participation of single sex schools. In comprehensive schools the boys' mean Calibration Test total score is higher (m = 35.4 & f = 32.3) than that for girls. This difference is statistically significant.

It should be noted that the male superiority in Calibration Test scores does not match the pattern of UK examination results at age 16, where girls outperform boys in many subjects (Stobart, Elwood and Quinlan, 1992). This pattern holds in the examination results of the candidates involved here, as will be shown later: confirming the need to be alert to gender issues when using the reference test in comparability studies. Clearly boys' Test scores must have a different relationship to their examination grades than girls'.

Given the existence of sex differences in performance on both the Calibration Test and the GCSE examinations, variations in correlational structure may be present which must be considered within the analyses of the data. Figure 2 illustrates this point.

Figure 2 Regression (locally weighted) of 1997 GCSE Maths 1661 grades on Calibration Test total scores for males and females



It shows the bivariate scatter of grades in a GCSE Mathematics examination (MEG Syllabus 1661) and Calibration Test total scores for males and females, together with the locally weighted least squares regression lines for grades on test scores for each sex. These (non-linear) regression lines thus represent the average grades achieved by males and females with given test totals and show how girls of a given level of general ability (obtaining a given Calibration Test score) outperform their male counterparts in GCSE Maths; except for the ablest boys and girls, who achieve quite similar average grades. This pattern is found in many other subjects, although the relative superiority of girls in some other subjects is

greater than in mathematics, as will be evident in subsequent analyses. The need to control for gender as well as general ability in applications involving the Calibration Test is thus very clear.

The non-linear regressions in figure 2 show little departure from linearity, suggesting that linear models will prove adequate for these data. Inspection of similar data showing relationships between Calibration Test scores and grades in a wide range of other GCSE syllabuses suggest that that an essentially linear relationship holds even where correlation is weaker and/or there is a greater gap between male and female performance.

School Type

Within the GCSE group, the scores of candidates from the twenty-three maintained comprehensive schools (n 2,307) have a markedly lower mean (33.8) and greater variance than those from the fifteen independent (44.0, n 971) and five maintained selective (45.7, n 378) schools involved, as would be anticipated. The comprehensive schools involved were drawn from a variety of regions and from rural, suburban and inner-city areas but did include three single sex girls schools. This feature apart, the comprehensive schools may well be a reasonably 'typical' selection, although the design of the 'sampling' and data gathering strategy were not primarily intended to assure this. It is more difficult to estimate how typical these particular independent schools (five boys schools; six girls schools and four mixed schools) and selective schools (two girls; one boys; two mixed) may be, given the varied character of schools in these sectors.

The differences in Calibration Test mean scores do not in themselves mean that the relationship between scores and examination grades varies systematically between pupils from different types of school, but this possibility needs to be examined.

Figure 3 illustrates this in the case of GCSE Mathematics 1661, showing the bivariate distributions of scores on the Calibration Test and the grades achieved by GCSE candidates from comprehensive schools on the one hand and the grades of those from independent or selective schools on the other. The selected nature of the candidates from independent/selective schools is clear, as is the tendency for pupils of moderate ability at such schools to obtain higher Maths grades than pupils of comprehensive schools with equivalent Calibration Test scores. This latter feature is summarised by the linear regressions of grades on Calibration Test scores plotted for pupils from both types of school. Whilst the very ablest pupils achieve similar grades irrespective of school type, the gap between the average grades of those from Independent/Selective[9] schools and those from comprehensives widens as ability (as measured by the Calibration Test) declines. This pattern prevailed in all subjects investigated in this study, as subsequent analyses will confirm.

There could be many explanations for this disparity. For instance varying levels of expectation by teachers, parents and/or pupils; or differentials in motivation, behaviour, study habits or family circumstances; or variations in resources or teaching or lower class sizes (Massey (1997) reported markedly lower GCSE Mathematics class sizes for pupils in the middle of the ability range in independent schools). Unfortunately this study can only note such variations in the relationship between Calibration Test scores and grades and can shed no light on which (or other) explanations make a difference. But clearly school type is a key factor which must also be taken into account when using general ability reference tests like the Calibration Test in investigations of equivalence of grading standards. The balance

[9] Which are similar in this respect, as separate analyses, not reported here, have confirmed.

between entries from different types of school for any given examination will affect 'expectations' concerning the distribution of grades which might be awarded.

Figure 3 Regression of GCSE Maths 1661 Grades on Calibration Test by School Type

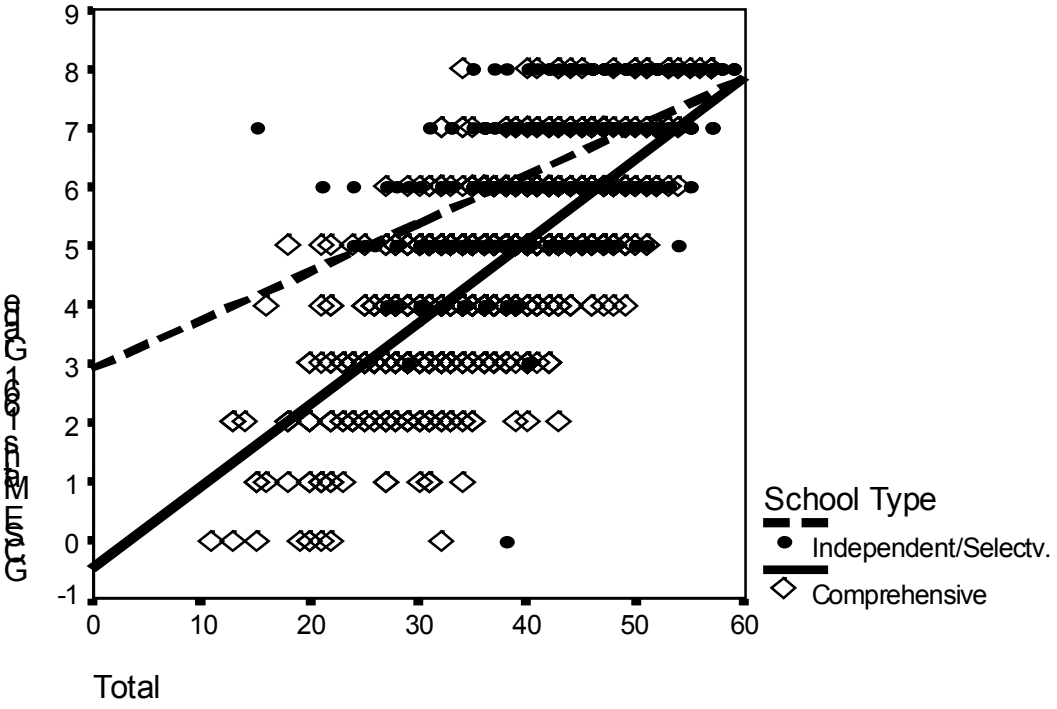


Figure 4 Regression of IGCSE/GCSE Maths Grade on Calibration Test by School Type

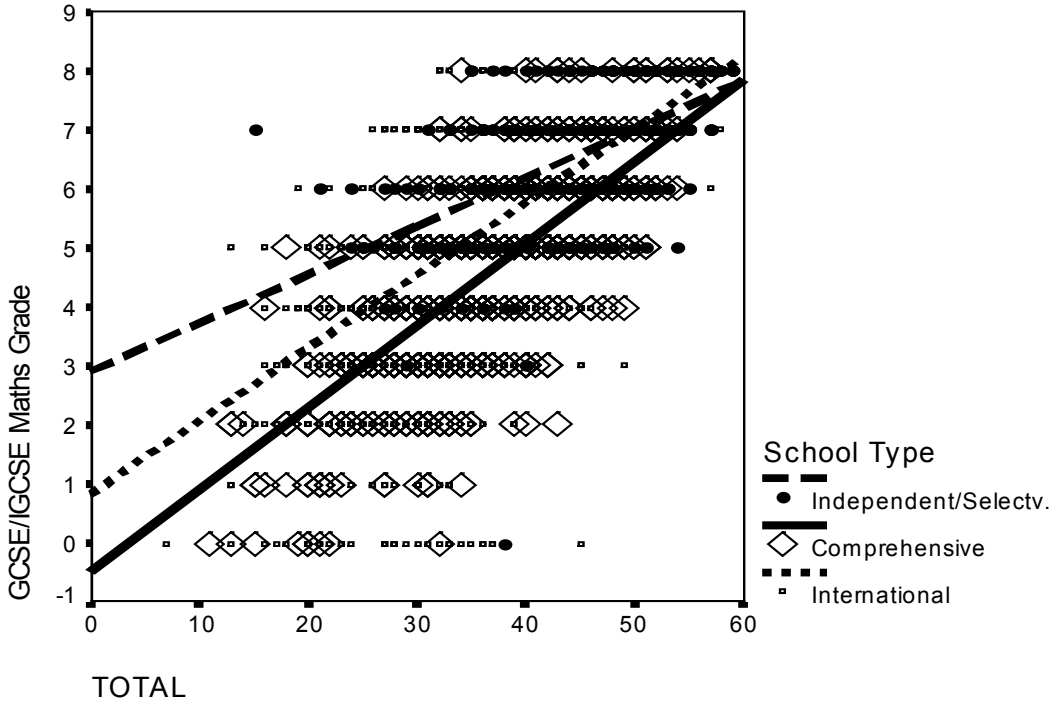


Figure 4 adds the equivalent data for IGCSE Mathematics 0580 candidates from international schools to the GCSE data presented in Figure 3. These IGCSE candidates are

drawn from a wider ability range than those from UK independent/selective schools and for a given level of ability (Calibration Test score) they obtain, on average, grades which fall between those awarded to pupils from UK independent/selective schools and comprehensives. Such comparisons are central to this study and in this instance would seem to suggest that it would be difficult to argue that IGCSE grading standards are out of line with UK GCSE practice, provided other factors (e.g. gender) do not disturb this portrayal of the data).

Home Language

The scores obtained by the relatively small sub-group of GCSE candidates (n 101) who do not normally speak English at home also differ from the scores of those who do. Non-EHL speakers' Calibration Test total scores were slightly lower (with a mean of 34.3 compared to 37.9) and somewhat more widely spread (sd 11.3 compared to 9.7) than those of EHL speakers, with their distribution including a 'tail' of low scores, as may perhaps have been expected. In contrast, in the IGCSE group the EHL (n 946) and (comparatively large) group of Non-EHL (n 639) candidates obtained the same mean Calibration Test score, with the scores of Non-EHL candidates the less widely spread. The scores of EHL and Non-EHL speakers on verbal and numeric or spatial reasoning items are shown in table 3, revealing that whilst UK GCSE Non-EHL candidates do worse than EHL candidates on both types of item, IGCSE Non-EHL candidates make up for their relatively poor performance on verbal items by outperforming EHL candidates on the non-verbal items.

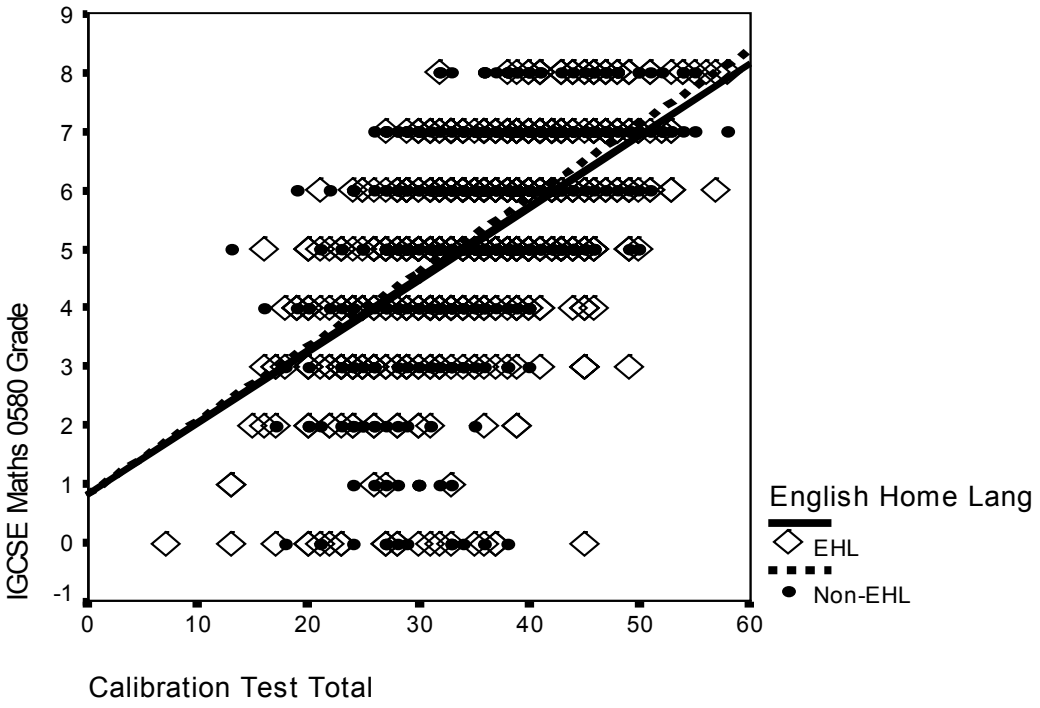
Table 3 Home Language and scores on items of different types

		n	*verbal*		*non-verbal*	
			mean	*sd*	*mean*	*sd*
GCSE	*EHL*	3,523	18.5	5.4	19.4	5.0
	Non-EHL	101	16.5	5.9	17.8	5.9
IGCSE	*EHL*	946	17.9	4.6	18.8	4.6
	Non-EHL	639	17.4	4.0	19.3	4.6

It thus seems likely that not speaking English at home is, in general, a qualitatively different matter in the IGCSE sample (where many pupils are members of non-English speaking communities and may have parents of different nationalities but have opted for an English medium education) than for the minority of Non-EHL pupils in UK schools. Accordingly it does not seem appropriate to treat this variable as meaning the same thing in the two examination groups in the course of the subsequent analyses. The few Non-EHL GCSE candidates will have no real impact on the regression of grades on test scores, but the language background of IGCSE pupils may affect outcomes in some subjects and thus requires consideration. Figure 5 investigates this in the case of IGCSE Mathematics 0580, showing the bivariate distribution of grades and Calibration Test scores and the regressions of grades on scores for the EHL and Non-EHL candidates respectively.

In this case it would seem that for a given level of ability (Calibration Test score) EHL and Non-EHL candidates obtain very similar average Mathematics grades, with the Non-EHL group achieving only slightly higher Mathematics grades, as their relatively high score on the non-verbal items might perhaps lead us to expect. But it seems unlikely that this pattern will hold for all subjects and whilst the Home Language of IGCSE candidates seems unlikely to be a major influence on grades, there may still be some merit in taking this factor into account whilst investigating comparability of standards.

Figure 5 Regression of IGCSE Maths 0580 grades on Calibration Test by Home Language



Analytic Models and Results

Hierarchical multi-level modelling techniques (Raudenbush & Bryk, 1986; Goldstein, 1987) were used to analyse the data, enabling the analyses to take the nested form of the data (in the form of school effects) into account, whilst simultaneously evaluating the effects of the different examinations and considering other factors.

Results are presented below from the use of two models to make comparisons between grading standards in the IGCSE and GCSE examinations set in a range of subjects.

Model 1

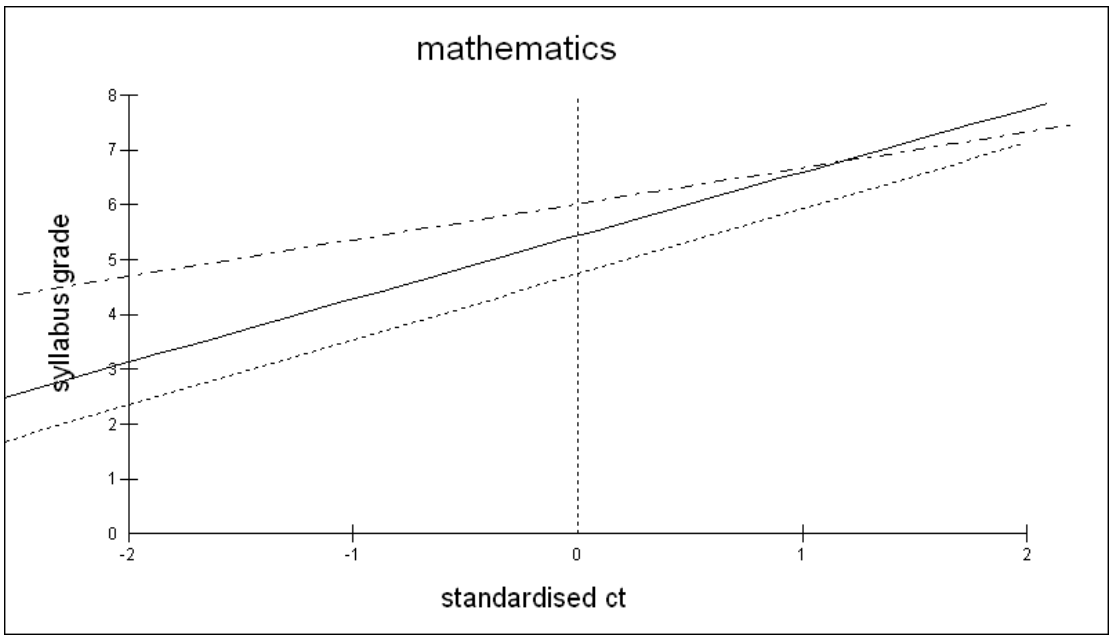
Model 1, given in Figure 6 and illustrated graphically in Figure 7, is a relatively simple model which estimates the variations in GCSE and IGCSE grades predicted for candidates of average ability (i.e. at the grand mean for Calibration Test scores of 38.4; these having been standardised prior to the analyses) from their Calibration Test scores, taking school type into account. The results for a series of comparisons of examinations in different subjects, presented in Table 4, provide the difference between the grades predicted for GCSE candidates from comprehensives and independent/selective schools respectively and for those for IGCSE candidates (i.e. the difference between their intercepts and that of the IGCSE candidates in Figure 7), together with the differences between the slope of the regressions of grades on test scores for the GCSE candidates from the two school types and the slope of the regression for IGCSE candidates, thus showing if the regressions are parallel. These enable us to test the null hypothesis that there are no significant differences between the grades, in a given subject, obtained by IGCSE candidates and those obtained by GCSE candidates from either comprehensives or from independent/selective schools.

Figure 6 Model 1

$$y_{ij} = \beta_{0ij}x_0 + \beta_{1j}x_{1ij} + \beta_2 x_{2j} + \beta_3 x_{3j} + \beta_4 x_{4ij} + \beta_5 x_{5ij}$$
$$\beta_{0ij} = \beta_0 + u_{0j} + e_{0ij}$$
$$\beta_{1j} = \beta_1 + u_{1j}$$

where
$_0$ refers to constant
$_1$ refers to standardised calibration test score
$_2$ refers to GCSE comprehensive pupils
$_3$ refers to GCSE independent/selective pupils
$_4$ refers to the interaction of x_1 & x_2
$_5$ refers to the interaction of x_1 & x_3

Figure 7 Model 1: graphical representation



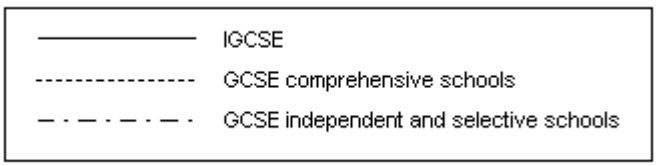


Table 4 Model 1 effect sizes:

Variation in GCSE and IGCSE grades (at the mean calibration test score) predicted from examination, calibration test scores and GCSE school type (comprehensive & independent/selective)

Syllabuses compared		GCSE comp - IGCSE		GCSE ind/sel - IGCSE		n		
GCSE	IGCSE	intercept	slope	intercept	slope	GCSE (comp)	GCSE (ind/sel)	IGCSE
Maths 1661	Maths 0580	-0.70**	+0.04	+0.58*	-0.50**	888	510	1243
English 1510	Eng 1st 0500	-0.32	+0.14*	+0.81**	-0.37**	1316	954	689
English 1510	Eng 2nd 0510	-0.28	+0.28**	+0.86**	-0.23*	1316	954	477
English Lit 1512	Lit B English 0486	+0.20	+0.10	+0.91**	-0.15	546	635	599
Science 1774	Biology 0610	-0.83**	-0.12	-0.04	-0.77	1143	14	1017
Biology 1780	Biology 0610	-0.06	-0.75	+0.62**	-0.81**	8	636	1017
Science 1774	Chemistry 0620	-0.80**	+0.12	-0.01	-0.53	1143	14	960
Chemistry 1781	Chemistry 0620	+0.44	-0.73	+0.55*	-0.46**	8	632	960
Science 1774	Physics 0625	-0.86**	-0.12	-0.07	-0.76	1143	14	878
Physics 1782	Physics 0625	-0.36	-0.39	+0.43*	-0.63**	8	585	878
French 1525	French 0520	-0.87**	+0.01	+1.00**	-0.33*	1007	919	422
Geography 1588	Geography 0460	-0.35	+0.18	+1.28**	-0.37*	418	168	702
History 1607	History 0470	-0.52*	-0.09	+1.06**	-0.69**	433	525	531
Art 1300	Art 0400	-0.68*	+0.20	+0.68	-0.19	414	177	278
IT Systems 1456	Comp St 0420	-1.50*	-0.12	+1.21	-0.68*	89	47	489
Bus St 1351	Bus St 0450	-0.35	+0.39	+0.08	-0.24	89	31	273
Economics 1485	Economics 0455	-0.16	-0.78	+0.06	-0.28	19	24	405

* statistically significant at 0.05 level
** statistically significant at 0.01 level

Thus, for instance, the results in table 4 relating to the comparison between the GCSE Mathematics 1661 and IGCSE Mathematics 0580 examinations confirm the display and summary of these data shown in figure 3 and the graphical illustration of the modelling in figure 5. The predicted grade at the intercept (with the average Calibration Test score) for comprehensive school candidates is 0.70 grades lower than that for IGCSE candidates, whilst that for independent/selective schools candidates is 0.58 grades higher. The slope of the regression predictions for comprehensive pupils is very similar to that for IGCSE candidates but the slope of the predictions for independent/selective pupils is significantly different, reflecting their flatter regression line, which converges towards the top of the ability range and indicates that weaker candidates in independent/selective schools fare relatively well in terms of grades. So although the IGCSE candidates may have obtained grades which are (statistically) significantly higher than those obtained by 'equivalent' comprehensive pupils in GCSE, it would be entirely unreasonable to suggest IGCSE grading standards are lenient, because the grades awarded to equivalent GCSE pupils from independent/selective schools are higher still.

Similar comparisons in other subjects yielded remarkably similar results. Those involving syllabuses in English Language (where GCSE English is compared with the IGCSE examination designed for candidates who speak English as a second language as well as the (more directly equivalent) First Language English examination), French, Geography, History, Art, Information Technology, Business Studies and Economics are straightforward and all follow essentially the same pattern: with the regressions for IGCSE grades on Calibration Test scores falling between the regressions for GCSE comprehensive school pupils (who, for any given ability level, obtain worse average grades than IGCSE candidates) and the regressions for pupils from independent/selective schools (who obtain higher average grades than IGCSE candidates).

The comparisons involving science syllabuses are less straightforward because the data has gaps which reflect UK curricular preferences, whereby few comprehensive schools enter pupils for the separate science examinations (Biology, Chemistry and Physics) and instead enter most of their pupils for the 'double' science syllabuses covering all three. International schools however prefer the separate sciences. The small numbers involved thus mean that it is not meaningful to compare grades awarded to comprehensive pupils in the separate sciences with those awarded to IGCSE pupils, although comparisons of their grades in (double) Science 1774 with all three separate IGCSE sciences fit the pattern established in other subjects (i.e. at a given ability level the IGCSE pupils of a obtain higher grades, on average). Similarly there are too few independent/selective school candidates for (double) Science to permit meaningful comparison but the grades of candidates from such schools for Biology, Chemistry and Physics are, on average, higher than those of IGCSE pupils of equivalent ability; again conforming to the pattern in other subjects. Here too, therefore, there seems to be no reason to suggest that IGCSE's grading standards are awry.

Only one subject fails to conform to this trend, English Literature 0486, where IGCSE candidates obtain average grades slightly lower than those of GCSE candidates from comprehensives as well as those from independent/selective schools. However the difference is not significant, falling within the margin of statistical error, and even here there seems no reason to suggest that there is any disparity in grading standards.

This simple model would therefore seem to suggest that IGCSE's grading standards are well in line with those pertaining in GCSE examinations across the range of disciplines. But we have earlier pointed out that school type is not the only variable which may need to be taken into account in making such comparisons. Accordingly a more complex model was fitted to the data.

Model 2

Model 2 tests the same null hypotheses as Model 1, but is a more complex model which has additional terms to estimate the effects relating to gender and the Home Language of IGCSE candidates, which we have already shown to have the potential to interact with the examination effects which are the focus of this study. The model is shown in figure 8 (although regrettably it is too complex to illustrate graphically) and results for comparisons in different subjects are given in table 5.

Figure 8 Model 2

$$y_{ij} = \beta_{0ij}x_0 + \beta_1 x_{1ij} + \beta_2 x_{2j} + \beta_3 x_{3j} + \beta_4 x_{4ij} + \beta_5 x_{5ij} + \beta_6 x_{6ij} + \beta_7 x_{7ij} + \beta_8 x_{8ij} + \beta_9 x_{9ij} + \beta_{10} x_{10ij} + \beta_{11} x_{11ij} + \beta_{12} x_{12ij} + \beta_{13} x_{13ij} + \beta_{14} x_{14ij} + \beta_{15} x_{15ij}$$

$$\beta_{0ij} = \beta_0 + u_{0j} + e_{0ij}$$

$$\beta_{1j} = \beta_1 + u_{1j}$$

where $_0$ to $_5$ are as model 1 and
$_6$ refers to IGCSE males
$_7$ refers to GCSE (comprehensive) males
$_8$ refers to GCSE (independent/selective) males
$_9$ refers to interaction of x_1 & x_6
$_{10}$ refers to interaction of x_1 & x_7
$_{11}$ refers to interaction of x_1 & x_8
$_{12}$ refers to non-English home language (IGCSE only)
$_{13}$ refers to interaction of x_1 & x_{12}
$_{14}$ refers to interaction of x_6 & x_{12}
$_{15}$ refers to interaction of x_1 & x_{14}

Table 5 Model 2 effect sizes:

(a) Examination effects: Predicted grades for female GCSE candidates from comprehensive and from independent/selective schools compared with those for female IGCSE English Home Language (EHL) candidates

Syllabuses Compared		Examination Effects (for females)				n		
GCSE	IGCSE	GCSE comp- IGCSE EHL		GCSE ind/sel -IGCSE EHL		GCSE comp females	GCSE ind/sel females	IGCSE EHL females
		intercept	slope	intercept	slope			
Maths 1661	Maths 0580	-0.59*	-0.08	+0.71*	-0.60**	470	267	350
English 1510	Eng 1st 0500	-0.58**	+0.04	+0.33	-0.46**	692	464	204
English 1510	Eng 2nd 0510	-0.53*	+0.06	+0.38	-0.45**	692	464	101
English Lit 1512	Lit B English 0486	+0.13	+0.08	+0.60*	-0.19	312	442	222
Science 1774	Biology 0610	-0.91	-0.16	-0.22	-0.87	618	0	348
Biology 1780	Biology 0610	-0.21	-0.16	+0.40	-0.89**	4	430	348
Science 1774	Chemistry 0620	-0.82**	+0.19	-0.12	-0.52	618	0	283
Chemistry 1781	Chemistry 0620	+0.15	-0.66	+0.33	-0.40*	4	427	283
Science 1774	Physics 0625	-0.73**	-0.09	-0.02	-0.78	618	0	219
Physics 1782	Physics 0625	-0.51	-0.51	+0.45*	-0.61	4	418	219
French 1525	French 0520	-0.80*	-0.14	+0.92**	-0.48*	512	585	163
Geography 1588	Geography 0460	-0.15	+0.32*	+1.13*	-0.48*	187	160	207
History 1607	History 0470	-0.32	+0.04	+1.05**	-0.76**	213	227	188
Art 1300	Art 0400	-0.55	+0.08	+0.72	-0.26	229	135	87
IT Systems 1456	Comp St 0420	-1.25*	-0.38	+0.99	-0.97*	34	34	111
Bus St 1351	Bus St 0450	-0.37	+0.25	-0.03	-0.52	56	23	83
Economics 1485	Economics 0455	-0.45	-5.92	+0.44	-1.20*	3	11	96

(b) Gender effects: Predicted grades for boys compared to those for girls in comprehensive schools, independent/selective schools and IGCSE schools

Syllabuses Compared		Gender Effects (m-f)			n		
GCSE	IGCSE	GCSE comp	GCSE ind/sel	IGCSE EHL	GCSE comp males	GCSE ind/sel males	IGCSE EHL males
Maths 1661	Maths 0580	-0.25**	-0.29	-0.15	418	243	349
English 1510	Eng 1st 0500	-0.61**	-0.14	-0.59**	624	490	209
English 1510	Eng 2nd 0510	-0.62**	-0.14	-0.54**	624	490	110
English Lit 1512	Lit B English 0486	-0.56**	-0.06	-0.50**	246	193	185
Science 1774	Biology 0610	-0.16		-0.33*	525	14	278
Biology 1780	Biology 0610	+0.27	+0.17	-0.33**	4	206	278
Science 1774	Chemistry 0620	-0.16		-0.32**	525	14	306
Chemistry 1781	Chemistry 0620	+0.73	+0.36	-0.32**	4	205	306
Science 1774	Physics 0625	-0.17		-0.01	525	14	308
Physics 1782	Physics 0625	+0.72	+0.23	+0.00	4	167	308
French 1525	French 0520	-0.74**	-0.47*	-0.99**	495	334	125
Geography 1588	Geography 0460	-0.72**	-0.24	-0.13	231	8	231
History 1607	History 0470	-0.75**	-0.32	-0.09	220	298	144
Art 1300	Art 0400	-0.50**[10]	-0.58	-0.15	185	42	77
IT Systems 1456	Comp St 0420	-0.35	+0.73	+0.11	55	13	163
Bus St 1351	Bus St 0450	-0.29	-0.10	-0.34	33	8	93
Economics 1485	Economics 0455	-0.10	-1.04[11]	-0.30	16	13	124

(c) IGCSE Home Language effects: Predicted grades for Non-EHL female candidates compared to EHL females and additional effect comparing male Non-EHL candidates with female Non-EHL candidates

Syllabuses Compared		IGCSE Home Language Effect		n	
GCSE	IGCSE	NonEHL-EHL (for females)	Gender m-f (NonEHL)	IGCSE NonEHL females	IGCSE NonEHL males
Maths 1661	Maths 0580	+0.26*	-0.04	200	290
English 1510	Eng 1st 0500	-0.52**	-0.01	118	141
English 1510	Eng 2nd 0510	-0.28	-0.05	105	138
English Lit 1512	Lit B English 0486	-0.18	-0.28	78	96
Science 1774	Biology 0610	+0.12	-0.14	167	182
Biology 1780	Biology 0610	+0.13	-0.13	167	182
Science 1774	Chemistry 0620	+0.22	-0.04	143	190
Chemistry 1781	Chemistry 0620	+0.22	-0.03	143	190
Science 1774	Physics 0625	+0.32*[12]	-0.19	116	201
Physics 1782	Physics 0625	+0.32*[13]	-0.19	116	201
French 1525	French 0520	+0.40*	+0.42	68	55
Geography 1588	Geography 0460	-0.09	-0.21	83	160
History 1607	History 0470	-0.07[14]	-0.35	85	95
Art 1300	Art 0400	+0.27	-0.49	44	63
IT Systems 1456	Comp St 0420	+0.56*	-0.68*	65	124
Bus St 1351	Bus St 0450	+0.35	-0.38	46	41
Economics 1485	Economics 0455	+0.25	-0.22[15]	62	104

* statistically significant at 0.05 level, ** statistically significant at 0.01 level

[10] * slope difference of -0.28
[11] * slope difference of +1.62
[12] * slope difference of -0.33
[13] * slope difference of -0.33
[14] * slope difference of +0.49
[15] * slope difference of +0.58

Do the results obtained with this more complex model revise the conclusions reached with model 1 above? The simple answer is no: the estimates of the effect sizes are modified but the conclusions remain the same in all subjects. The pattern of differences between the IGCSE and GCSE examinations in all subjects remains very much the same when gender and home language have also been taken into account, as the figures in section (a) of table 5 reveal. There thus seems no reason to doubt that IGCSE grading standards conform to those established in the GCSE examinations.

There are of course quite substantial differences between the performance of male and female candidates. Section (b) shows that in this sample of IGCSE candidates, EHL girls (i.e. those who speak English at home) tend to obtain better grades than EHL boys in many subjects. The difference (in average grades) varies, being about one whole grade in French, about half a grade in English Language (First & Second) and Literature, and about one third of a grade in Biology and Chemistry. These differences are all statistically significant. Substantial differences are also apparent in Business Studies and Economics, but these fail to achieve statistical significance because of the relatively small numbers of pupils concerned. However the IGCSE EHL girls' superiority is less marked in the humanities and Mathematics (only about one tenth of a grade, well within the margin for sampling error). Notable exceptions are Physics (where both sexes obtain the same average grade) and Computer Studies (where IGCSE boys outperform girls slightly), though again the differences are not significant. Much the same pattern of sex differences is present in the grades awarded to Non-EHL IGCSE candidates (except in French -see below). Similar sex differences are also evident in the performance of GCSE candidates in both comprehensives and independent/selective schools, although the effect sizes vary somewhat, with UK comprehensive school girls' superiority in the humanities and art proving more substantial than is evident in IGCSE.

The language candidates' normally speak at home is also relevant to the grades we might expect them to achieve in different subjects, although the effects of home language appear to vary in different subjects and may be different for males and females in some cases.

Given that these candidates have been taught and examined in English, one might naively expect EHL candidates to obtain higher grades than Non-EHL candidates with equivalent Calibration Test scores, but in fact this would be mistaken and is only case in the cluster of English subjects, where we should note that the entry patterns of candidates with different language backgrounds are not as clear-cut as the syllabus nomenclature seems to suggest. In First Language English[16], girls and boys with an EHL background achieve results on average about half a grade better than those whose home language is not English who have similar Calibration Test scores (a statistically significant gap). In English 0510[17], which is designed for those who are not native speakers, the pattern is similar, although the slope of the regression of grades on Test scores is less steep for those (both boys and girls) whose home language is not English and the gap between candidates from different linguistic backgrounds does not reach statistical significance in this group. In English Literature[18] there seems to be relatively little difference in the average grades of female candidates from EHL and Non-EHL backgrounds, controlling for Calibration Test scores, and although

[16] Where this sample included 413 EHL candidates and 259 Non-EHL candidates.
[17] Where this sample included 243 Non-EHL candidates and 211 EHL candidates.
[18] Where this sample included 407 EHL candidates and 174 Non-EHL candidates.

Non-EHL boys tend to do less well than their EHL equivalents in this subject, their superiority is also not sufficient to be regarded as statistically significant. In both Geography and History it is again the case that whilst EHL candidates do obtain higher average grades, such differences are relatively small and not statistically significant, although here again the Non-EHL boys seem to achieve disappointing grades by comparison with others with equivalent scores on the Calibration Test.

In other subjects the position is reversed, with Non-EHL pupils at a given Calibration Test score being likely to obtain higher grades than their EHL equivalents. This trend proves strong enough to reach statistical significance in Mathematics, Physics, Computer Studies and French.

The fate of Non-EHL boys in French is in marked contrast to that of such boys in other subjects, perhaps reflecting an advantage gained by some at least from their linguistic background. In French Non-EHL boys tend not to show the performance deficit exhibited by their English speaking counterparts, with abler candidates especially being likely to obtain grades similar to those of girls - even the excellent grades obtained by Non-EHL girls.

Discussion and Conclusions

We should perhaps reflect upon the meaning of the home language variable. Non-EHL pupils have said that they do not normally speak English at home but, this acknowledged, they may or may not have parents who are native English speakers and their exposure to English speaking environments can vary extensively. Conversely, candidates who report that they do normally speak English at home may live in families where there are no native English speakers. Without doubting the accuracy of the candidates' questionnaire responses, these EHL and Non-EHL categories may mean very different things for different people. The linguistic experience, at home or at school, of the groups of EHL candidates entering for First and Second Language English may be quite different, and may reflect appropriate judgements about their language experience, although it is impossible for us to say how or how far this is true. It is moreover difficult to see how to gather better explanatory information, and the explanatory variable we have is better than none at all.

We should also reflect upon the methodological implications of language background. Should we expect Non-EHL candidates to obtain similar grades to EHL candidates who have similar Calibration Test scores? Was the common yardstick we have used an equally 'fair' monitor for both these groups and how is the null hypothesis we have tested affected?

The answer to the first of these questions is patently no; a conclusion supported by both logical reasoning and the empirical evidence. Empirically we can see that the relationship between Calibration Test scores and grades for EHL and Non-EHL candidates is not always the same. In English Language EHL candidates at a given Calibration Test score obtain higher grades (on average) than Non-EHL candidates, whilst this position is reversed in Mathematics and some other subjects. But we should expect this to be the case. It would be unrealistic to expect the Calibration Test to be free of cultural or linguistic or other biases and we must consider the possibility and implications of these before reaching any conclusions which depend upon a reference test as the basis for comparisons. Clearly we must acknowledge that a group of non-English speaking candidates may be at some disadvantage in taking the Calibration Test, even where they have been educated in English. Given this, Non-EHL candidates may need to be somewhat more able than EHL candidates to obtain the same Calibration Test score and, being more able, might thus be expected to

obtain higher examination grades than EHL candidates in those subjects where language skills provide little or no direct advantage, although the reverse might be expected in subjects where language skills are at a premium. This is exactly the pattern we have observed empirically. Non-EHL candidates do better in Maths etc. but worse in English and some may even benefit from a linguistic advantage in French!

But we have explicitly recognised and controlled for such potential for bias in calibration Test scores with respect to groups of candidates (including those of different gender and from different types of school as well as those from different language backgrounds) in the research design and in the model for data analysis. The effects of home language (in so far as we have been successful in categorising this) have been distinguished from effects relating to the IGCSE and GCSE examinations via comparisons for EHL and Non-EHL candidates respectively.

Nothing in the discussion above implies that the Calibration Test is biased or unequally relevant with respect to the examinations (i.e. IGCSE subject x versus GCSE subject x) being compared, nor does it negate our efforts to make such comparisons. Furthermore, the possibility that the Calibration Test might be unequally relevant with respect to the examinations compared is itself investigated within the models for data analysis, which explore the parallelism of the regression of grades on Test scores for candidates taking different examinations, as well as the significance of any differences in intercept.

Thus all that remains to establish the validity of the assumption that the Calibration Test is equally fair and relevant (one of the two major assumptions in the operational definition of equivalent standards) is the possibility that it might be biased with respect to one or other examination. For instance, would we expect that following an IGCSE Geography course confers any advantage or disadvantage, with respect to scoring on the Calibration Test, by contrast with the GCSE Geography syllabus? This seems very unlikely, given the Calibration Test's nature, deliberately distanced from the subject-based curriculum.

Nor does the possibility that one syllabus/group might prove more fertile ground for converting ability into achievement (the definition's other main assumption) give rise to concern. The use of school type as a control variable also goes some way to reassure on this front, enabling comparisons with their different curricular (in its widest sense) settings. Simplistic comparisons have been avoided and the way in which the data for international schools appears to occupy the intervening ground between the two types of UK school is reassuring. Hence we can be reasonably sure that the methods we have used for drawing conclusions about respective examination standards in GCSE and IGCSE are soundly based.

These issues will however be relevant to any future efforts to use the Calibration Test as a vehicle for comparing grading standards. It will be important to recognise the effects of factors such as school type and gender and language background when establishing the base against which comparisons should be made. So to compare GCSE standards at some future date with those pertaining in 1997 it would be necessary to control for gender and school type as well as general ability. Likewise comparisons between June 1997 IGCSE standards and IGCSE standards at any other time will need to control for language background. Whatever the deficiencies of this variable, it is sufficiently important that the relationships established here between IGCSE grades and Calibration Test scores for Non-EHL girls and boys may provide the most appropriate baselines available, should we in future wish to make comparisons of grading standards with other examinations where many of the candidates are from a Non-EHL background. Given the main substantive conclusion

of this study, which is that there seems no reason to doubt that the IGCSE examiners have been successful in their efforts to implement GCSE grading standards, such comparisons should be valid.

This study has served its purpose in screening grading standards in IGCSE for disparities with the GCSE and no further action seems to be required at present, unless professional judgements themselves suggest that there are legitimate concerns which would be unlikely to have been detected by this method of comparison. It would however be desirable to repeat the exercise at some future date, say in three to six years time, when it should also be possible to explore the possibility that standards may have drifted over time by using the current data as a baseline. A data archive for applications of the Calibration Test to investigate standards will need to be maintained to facilitate this.

The differences in achievement relating to gender and home language revealed in the course of controlling for these variables have proved interesting in themselves and may warrant further investigation, which we hope to be able to pursue.

Summary

Grading standards in the June 1997 IGCSE examinations in fourteen subjects, spanning the full range of the curriculum, were contrasted with those in similar GCSE examinations. Comparisons used a 'Calibration Test' (made up of items testing verbal, numerical and spatial reasoning) to control for variations in general ability between candidates for the different examinations.

The data confirmed the further need to control for variations between the groups of candidates taking the different examinations with respect to gender, the type of school attended by UK GCSE candidates and the language background of IGCSE candidates. All of these factors (and interactions between them in some cases) proved to affect the relationship between grades obtained and Calibration Test performance. Nevertheless the research design and multi-level models employed in the analysis of the data appear to prove adequate.

On average, IGCSE candidates obtained slightly higher grades than GCSE candidates of equivalent ability from comprehensive schools. However they obtained lower grades than GCSE candidates from independent/selective schools and there was no suggestion that IGCSE grading standards in any of the subjects investigated were more or less lenient than those established in the GCSE examinations.

Variations in performance relating to gender, school type and language background proved interesting in themselves and may merit further investigation.

References

Bardell, G.S., Forrest, G.M. & Shoesmith, D.J. (1978*) Comparability in GCE: A Review of the Boards' Studies 1964-1977*, Manchester, Joint Matriculation Board on behalf of the GCE Examining Boards.

Gipps, C. & Murphy, P. (1994) *A Fair Test?: Assessment, Achievement and Equity*, Buckingham, Open University Press.

Goldstein, H. (1987) *Multilevel Models in Educational and Social Research*, London, Griffiths.

Massey, A.J. (1997), 'Variations in class sizes and achievement in GCSE mathematics by pupils from LEA/GM and independent schools', *Research in Education, 57, 25-35*.

Massey, A.J., Bramley, T., Dexter, T. & McAlpine, M. (1998*) Calibration Test: Design, Development and Validation*, Cambridge, Research & Evaluation Division, UCLES.

Murphy, R.J.L. (1982) 'Sex differences in objective test performance'*, British Journal of Educational Psychology, 52, 213-219*.

Murphy, R., Wilmut, J. & Wood, R. (1996*) The Use of Reference Tests for Monitoring A Level Standards*, London, School Curriculum and Assessment Authority.

Newbould, C.A. & Massey, A.J. (1979) *Comparability using a Common Element*, Occasional Publication 7: Test Development and Research Unit, Cambridge, TDRU.

Nuttall, D.L. (1971) *The 1968 CSE Monitoring Experiment*. Schools Council Working Paper 34, , London, Evans/Methuen Educational.

Raudenbush, S.W. & Bryk, A.S. (1986) A hierarchical model for studying school effects, *Sociology of Education, 59, 1-7*.

Stobart, G., Elwood, J. & Quinlan, M. (1992) Gender bias in examinations: how equal are the opportunities? *British Educational Research Journal*, 18, 3, 261-76.

UCLES (1995) *IGCSE: an Introduction*, Cambridge, UCLES.

Willmott, A.S. (1997) *CSE and GCE Grading Standards: the 1973 Comparability Study*. Schools Council Research Study, London, Macmillan Education.

Willmott, A.S. (1980) *Twelve Years of Examination Research: ETRU 1965-1977*, London, Schools Council.

Wrigley, J., Sparrow, F.H. and Inglis, F.C. (1967*) Standards in CSE and GCE: English and Mathematics*. London, Schools Council Working Paper No 9: HMSO.