Predicting Final Grade Distribution of Examination Results – Selection of Sample Scripts

Hugh McManus, State Examinations Commission, Ireland.

Abstract

The State Examinations Commission is responsible for administering public examinations at the end of lower second-level and upper second-level education in Ireland. Unusually for such examinations, grade boundaries are predetermined and no transformations are applied to raw scores. The raw mark therefore directly determines the grade, and no manipulation of the grade distribution is possible after the raw marks are finalised. For security reasons, examination items are not pre-tested. When scripts are being marked, therefore, early and accurate predictions of the overall distribution are essential, so that any necessary interventions can be made at the raw-mark level, by adjusting marking criteria and revisiting scripts already marked. In order to generate this initial estimate of the final distribution, each examiner marks a sample set of scripts early in the process.

This paper addresses the mechanism used for selecting the sample. It does not deal with the issue of determining an appropriate sample size. It reports empirical confirmation that the mechanism currently recommended to examining teams (systematic sampling) produces better estimates than might be expected from random samples. It also confirms that certain alternative mechanisms previously used by some examining teams will produce biased samples in most circumstances.

1. The Certificate Examinations in Ireland

The **Leaving Certificate Examination** in Ireland marks the end of upper second-level education. Its stated purpose is to provide certification of achievement on the second-level curriculum. In reality, it also serves, in most instances, as the sole determiner of entry into universities and other third-level institutions. This is accomplished through a centralised application process that has been established co-operatively by those institutions. The Central Applications Office receives examination data directly from the State Examinations

Commission and determines a rank ordering of candidates using a simple composite score[1]. Places are awarded to applicants according to this rank, subject to minimum entry requirements for the course in question. In the vast majority of cases, there are no interviews, aptitude tests, or other supplementary criteria. The examination may therefore be considered to be a high-stakes one.

The **Junior Certificate Examination** marks the end of lower second-level education and of the period of compulsory schooling. The great majority of students continue their education in the same school, largely without restriction, so this examination serves little if any gate-keeping function. Nonetheless, it is currently modelled to a large extent on the Leaving Certificate.

In the case of both certificates, most subjects are available to be studied at two levels (higher and ordinary). Furthermore, examinations in the various subjects may consist of several components. Each examination has at least one terminal written component, and some examinations have other components, such as listening comprehension tests and oral tests in the languages, practical skills tests, project and portfolio work, etc. Almost without exception, examination components are examined externally. The examiners are teachers, but not the candidates' own teachers.

2. Marking and the Award of Grades

Predetermined grade boundaries are laid down by regulation, and no standardising transformations are permitted to be applied to raw scores. Hence, the final raw score awarded by the examiner directly determines the grade received by the candidate. For example, at leaving Certificate, a candidate is awarded an A1 grade if he or she achieves 90% of the available marks. These fixed boundaries obviously pose considerable difficulties for maintaining year-on-year comparability of results. It is accepted that some level of grade variation from year to year is inevitable, but, since there is usually no reason to believe that there are substantial shifts in the overall level of achievement in the cohort over short periods of time, large variations from established patterns are not deemed acceptable. It is assumed in such circumstances that the shift is more likely to be due to the effects of the particular test instrument and its application than to a sudden change in cohort achievement. Accordingly,

[1] CAO assigns a numerical score to each grade and calculates the total of the best six grades. There is no calibration or re-scaling – an A1 at higher level in one subject has the same value as an A1 at higher level in any other.

some intervention is appropriate to bring the distribution into closer alignment with the expected one. Since policy dictates that no *post hoc* transformation may be applied, the adjustment must be made to the raw marks on the scripts. Clearly, then, early detection of the emerging distribution is vital.

It should be noted that, in what follows, sampling issues are conceptualised as they relate to batches of written papers; the same issues arise to a greater or lesser degree with other types of component.

3. Gathering Sample Data

Before the marking of a particular component begins, the examiners attend a conference at which they are trained in the accurate application of the marking scheme. Following this, they begin their work by marking a sample of scripts from their batch. They transmit the resulting grades through the hierarchy of examiners to yield the overall picture. If the grade distribution on these sample scripts is satisfactory, marking proceeds. Otherwise, amendments to the marking scheme are made and are communicated to the examiners, who re-mark the sample scripts, report the revised statistics, and then continue.

In 2000, a working group looking at various aspects of the examining process considered the issues of both sample size and the mechanism for selecting the sample. This group concluded that the previously existing practice of selecting 20 scripts from each examiner's batch should continue in the majority of cases. In cases where this would lead to an overall sample size of less than 600, larger samples should be taken, subject to time and other logistical constraints.

Until that time, the officially preferred mechanism for selecting the sample was for the examiner to attempt to select randomly from throughout the batch. However, some examining teams used other procedures, and instructions given by individual managers varied. For example, examiners were sometimes instructed to take one script from each of the first twenty packets in the batch. The working group recognised the problems of the previous official position, including both the well-documented difficulties in relying on human judgment in selecting random samples, and the reality that not all teams were following the protocol. The group recommended switching to a (random-start) systematic sample[2] within each

[2] For example, for a sample of 20 scripts from the batch, the examiner calculates k = (batch size) / 20 and then selects every k^{th} script, starting with one selected at random from the first k–1 scripts.

examiner's batch. This was deemed easy for all examiners to implement, and likely to lead to a good sample. It was made the official position and is now almost universally implemented.

The current study arose from the fact that maintaining an appropriate distribution in each subject is still a significant focus of the management of the marking process, and there is a consequent need to ensure that the sampling procedures are as good as they can be in the circumstances. It was also noted that not all teams are following the current protocol, and it is useful to clarify what the consequences of that may be. Furthermore, it is useful for the managers of the process to know just how accurate the samples are proving to be, in order to make reasonable intervention decisions in the context of this and other indicators.

4. Characteristics of an Examiner's Batch

Scripts are packaged at the school by the superintendent of the examination centre and transmitted to the State Examinations Commission. Centres vary in size, and most schools will have a number of centres, usually numbered consecutively. At the Commission, these intact packets are arranged into examiner batches, largely on a county-by-county basis, such that each examiner receives scripts from counties remote from where he or she lives and teaches. This arrangement has a number of consequences for the characteristics of the batch:

1. The examiner's batch is not a random sample of the cohort as a whole, and is unlikely to be representative of the cohort. It will consist of scripts from a limited geographical area and will probably not be proportionally representative of various school types.
2. The scripts are not randomly distributed within the batch: scripts from candidates in the same school are generally adjacent, and within that, scripts from candidates with the same teacher are possibly adjacent. Furthermore, assignment of examination numbers within a centre may be systematically patterned (e.g. by surname or date of birth).

As an aside, the first matter above poses difficulties for establishing any meaningful quantitative criteria for determining whether an individual examiner's grade distribution is sufficiently out of line with norms to justify questioning the quality of that examiner's work. Class-level, school-level, and geographical factors are known to exist in candidate achievement, and this will cause the distributions from batch to batch to vary far more than would be the case with simple random samples of the same size.

The second matter above is the crucial one in the current context. Any sampling procedure that systematically favours certain scripts within the batch is a potential source of bias in the sample. Examples include procedures that favour scripts from smaller packets over scripts from larger packets (or vice versa), ones that favour early scripts within a packet over later scripts, ones that favour scripts near the start of the batch over those near the end, etc. Some of these potential sources of bias may not actually result in a biased outcome, but they are all worthy of consideration.

5. Statistical Sampling Methods

The advantages and disadvantages of various standard sampling methods are given in texts on the subject. Simple random sampling, although theoretically satisfactory, is in practice often difficult to implement accurately. In particular, the requirement that each member of the population must have an equal probability of selection poses a problem in many contexts, including the one under discussion. Stratified sampling, given suitable sampling strategies within each stratum, can provide more accurate estimates under certain relatively common circumstances. Nonetheless, to implement properly, it requires a considerable degree of effort on the part of the person selecting the sample, and also requires one to already have, or to simultaneously generate, statistical information at stratum level.

Systematic sampling has a number of advantages in this context. It is easy and fast for a non-expert to implement by hand. It avoids forms of bias inherent in human judgments of randomness. And, if there are indeed homogeneous subgroups in the population, then a systematic sample will usually result in what is effectively a proportionally sampled stratified sample. Hence, in many circumstances there is a good chance that the estimates will be better than those given by a simple random sample.

A potential difficulty with systematic sampling is that cyclical patterns in the sampling frame can cause a biased sample. However, there is no reason to expect such periodicities in examiner batches and none has been noted.

6. Analysing the Effectiveness of Existing Sampling Methods

Some *post hoc* analyses of the full batch of scripts in a range of subjects were undertaken. These looked at two sampling methods: "first script in each packet" sampling, and random-start systematic sampling.

6.1 Sampling by Taking the First Script in Each Packet

The first analysis undertaken was intended to check the outcome of a sampling method used by some examining teams: taking the first script from each of the first twenty packets. This might strike one as a peculiar sampling method to use in the first instance, but the reality is that some examining teams have used it in the past and may still be doing so. One possible reason for its use is that, in some cases, separate examining teams are marking the different components (e.g. Paper 1 and Paper 2). By implementing this sampling method, the same candidates are being sampled on each paper, and the teams are able to pair the sampled scripts to give a view of the grade distribution of the composite scores.

Since batch-level data were not readily available, this sampling method was modelled as a selection of the first script from every packet in the whole script population. It was suspected that the sampling procedure might result in a biased sample, and this proved to be indeed the case (see Table 1). The suspected reason for the bias was also confirmed: there is almost always a relatively small but statistically significant correlation between the size of the centre and candidate achievement. For example, in the case of Leaving Certificate Higher-Level Mathematics, the correlation is 0.14. That is, candidates from larger centres score, on average, slightly better than candidates from smaller centres. Since the sampling procedure skews the sample in favour of scripts from smaller centres, the resulting sample significantly underestimates candidate performance. It therefore underestimates the A-rate and overestimates the failure rate. In this case, the A-rate in the population is 15.7% whereas the A-rate in the sample is 13.6%. A sample of this size should yield an estimate of the A-rate that is accurate to within 0.6

Statistics for sample consisting of first script from each packet (ATAL Maths, 2005)	
Population mean mark (out of 600)	406.89
Population standard deviation	97.82
Size of sample	1675
Standard error of mean for random sample of this size	2.39
Observed sample mean	397.45
Corresponding observed z-value	–3.7
Corresponding p-value	0.0001

Table 1: "First script in each centre" produces a biased sample (significantly underestimates performance on Higher-Level papers)

percentage points, so this is a highly significant underestimate, with the potential to lead to inappropriate intervention decisions.

Variants of the above procedure were also tested and produced similar results (e.g. last script in every centre, middle script in every centre, first script in every other centre). In this context it should be noted that systematic variations within centres were not detected. Achievement is not well correlated with relative position within the centre[3], and, correspondingly, estimates arising from taking the last script in each packet are similar to those arising from taking the first.

The statistics for English, Maths, and Geography at each level are presented in Table 2 in the appendix. It is interesting to note that the correlation between centre size and achievement is positive for Higher Level papers in each subject, and negative for Ordinary Level papers in English and Geography. The correlations are stronger at Higher Level. In the case of Ordinary Level Mathematics, the correlation is not statistically significant, nor is the error in the sample mean. The error in the A-rate is highly significant, however, indicating perhaps that centre size may be negatively correlated with achievement at the upper end of this cohort[4]. Reasons for this pattern of correlations have been considered but are not relevant to the issue at hand; all that concerns us here is that the correlations exist.

6.1 Random-Start Systematic Sampling

The second analysis undertaken involved checking lots of systematic samples, in order to test whether they are better, worse, or equally as good as those that would be expected from a simple random sample of the same size. Conceivably, depending on the subject, the sampling could vary between taking every tenth script and taking every twenty-fifth script. Accordingly, for each period from 10 to 25, all systematic samples of that period were checked, to yield observed sampling distributions of A-rates, failure rates, and mean scores. Bear in mind that, for example, the observed A-rate for a given sample is not independent of the observed failure rate, or the observed mean score, so it is safer to look at just one of these at a time when evaluating the quality of the sampling process. These observed sampling distributions were then compared to the theoretical sampling distributions that a simple random sample of the

[3] Statistically significant, but small, correlation was found at Ordinary Level in Maths and English. No significant correlation was found in other datasets.
[4] An analysis of this hypothesis was not undertaken.

same size would be expected to yield. In particular, the observed standard error of the statistic involved was compared with the theoretically expected standard error for random samples.

To illustrate, consider the mean mark on the examination, as estimated by a large systematic sample of period 15. There are 15 different such systematic samples. If these behaved like simple random samples, then the standard deviation of the 15 estimates would be expected to equal the theoretical standard error for random samples of the same size. This would similarly be the case for other periods of interest. When this was tested, it was found that in the great majority of cases, the "observed standard error" was smaller than the "theoretical standard error". When averaged over the different possible periodicities, the ratio of observed to theoretical standard error was less than 1 for all datasets analysed, (Table 3 in the appendix). It varied from 0.86 for Higher Level English to 0.99 for Ordinary Level Mathematics, with an average of 0.92 over the seven datasets.

It appears, then, that the suggestion made in section 5 above is supported: the systematic samples are giving somewhat more efficient estimates than might be expected.

7. Conclusions

The sampling method that SEC examining teams are currently directed to use (random-start systematic sampling) is highly satisfactory and departures from it should be discouraged. In particular, any deviations that introduce centre-dependant factors should be avoided.

Confidence intervals for estimates of various levels of achievement (such as A–rates) can afford to be slightly tighter than those calculated on the basis of simple random sampling theory. They can probably be reduced by 5–10%. Equivalently, if the target sample size is being chosen to achieve a confidence interval of a specified size, the sample can probably afford to be 10–20% smaller than might be expected.[5]

On a broader issue, despite the difficulties that the use of fixed grade boundaries imposes on maintaining grade comparability, there appears to be no appetite in Ireland to consider changing the practice. Having a system that is easily understood by the public is highly valued, especially since candidates have full access to their marked scripts. The current system has the

[5] The standard error is inversely proportional to the square of the sample size, so the ratio of the theoretically required sample size to the actually required sample size is the square of the ratio of the actual standard error to the theoretical one.

advantage that candidates understand that if their total mark reaches a pre-ordained known threshold, they will receive the corresponding grade. An unfortunate disadvantage is that a disproportionate amount of effort during the early part of the marking process is expended on ensuring an acceptable distribution. Such effort might be better directed at ensuring consistency and fairness in the application of the marking scheme. It must also be recognised that any amendments made to a marking scheme in order to adjust the distribution are likely to make the scheme somewhat less valid than would otherwise be the case. Changing to a system of establishing cut-scores after the marking has finished would render the monitoring and adjustment of the mark distribution unnecessary, and would facilitate the refocusing of efforts in more appropriate directions. Although a robust system for setting cut-scores may not be straightforward to achieve, such a move is at least worthy of serious consideration.

Appendix: Summary of Statistics

Subject	Population mean	Sample mean	Population A-rate	Sample A-rate	Population failure rate	Sample failure rate	Correlation between score and centre size	Correlation between score and position in centre
Higher Level Mathematics	406.27 (out of 600)	396.99 ($p = 0.00006$)	15.7%	13.0% ($p < 10^{-6}$)	5.2%	6.0% ($p < 10^{-6}$)	0.14 ($p < 10^{-6}$)	0.007 ($p = 0.5$)
Higher Level English	258.12 (out of 400)	251.56 ($p < 10^{-6}$)	10.3%	8.2% ($p < 10^{-6}$)	1.8%	3.0% ($p < 10^{-6}$)	0.15 ($p < 10^{-6}$)	0.002 ($p = 0.7$)
Higher Level Geography	254.53 (out of 400)	247.53 ($p < 10^{-6}$)	9.2%	7.4% ($p < 10^{-6}$)	4.4%	4.9% ($p < 10^{-6}$)	0.13 ($p < 10^{-6}$)	–0.004 ($p = 0.5$)
Ordinary Level Mathematics	376.56 (out of 600)	379.21 ($p = 0.3$)	13.3%	14.4% ($p = 0.00004$)	13.7%	13.6% ($p = 0.6$)	0.01 ($p = 0.2$)	0.02 ($p = 0.00004$)
Ordinary Level English	256.23 (out of 400)	260.81 ($p = 0.00004$)	8.0%	9.5% ($p < 10^{-6}$)	2.8%	2.2% ($p < 10^{-6}$)	–0.10 ($p < 10^{-6}$)	0.02 ($p = 0.002$)
Ordinary Level Geography	248.39 (out of 400)	251.81 ($p = 0.04$)	7.0%	8.5% ($p < 10^{-6}$)	7.3%	7.5% ($p = 0.3$)	–0.08 ($p < 10^{-6}$)	0.01 ($p = 0.34$)
Foundation Level Mathematics	382.48 (out of 600)	387.12 ($p = 0.07$)	7.6%	8.0% ($p = 0.03$)	8.9%	7.4% ($p < 10^{-6}$)	–0.03 ($p = 0.01$)	–0.003 ($p = 0.8$)

Table 2: Observed errors in samples based on first script in each packet.

Subject	$\frac{\text{mean observed S.E.}}{\text{S.E. of random sample}}$
Higher Level Mathematics	0.92
Higher Level English	0.86
Higher Level Geography	0.91
Ordinary Level Mathematics	0.99
Ordinary Level English	0.89
Ordinary Level Geography	0.97
Foundation Level Mathematics	0.88

Table 3: Comparison of observed standard errors of relevant systematic samples with standard errors of random samples of the same size.