A Hybrid Approach to Moderation of School-based Assessment in Advanced Supplementary Level Liberal Studies – Using Expert Judgement and Bayesian Hierarchical Statistical Modelling to Enhance Reliability and Comparability

Lo Ka-yiu
Hong Kong Examinations and Assessment Authority, Hong Kong
kylo@hkeaa.edu.hk

Fung Tze-ho
Hong Kong Examinations and Assessment Authority, Hong Kong
thfung@hkeaa.edu.hk

Abstract

In the Hong Kong Advanced Level Examination, Liberal Studies (LS) requires each student to complete an individual research project. As a mode of school-based assessment (SBA), the projects are marked by their teachers and counted as part of the public examination. However, teachers are not necessarily aware of the standards of performance across all schools. To achieve comparability of assessments across schools, the project marks awarded by teachers will be moderated statistically.

In 2010 LS exam, moderation will be conducted on school basis. In principle, 5 projects will be selected from each school from different levels of performance in order to obtain a representative sample. Each sampled project will be double marked by two external assessors. Based on external assessors' marks, the school performance level on SBA and the corresponding variability are estimated. However, the reliability of these statistics may be cast in doubt in view of small sample size. With respect to this problem, the use of Bayesian hierarchical statistical modeling is proposed so as to share information across different schools in order to increase the reliability of statistics concerned. Empirical study shows that the approach is promising in stabilizing the estimations and preventing excessive changes to teachers' marks.

Keywords: Liberal Studies, school-based assessment, expert judgment, statistical moderation, Bayesian hierarchical method.

Introduction and Background

As stipulated in the Hong Kong Advanced Supplementary Level Liberal Studies (LS) syllabus, there are totally 6 exam papers on the following modules, namely: (i) Human Relationships (HR); (ii) Hong Kong Studies (HKS); (iii) Environmental Studies (ES); (iv) China Today (CT) ; (v) Modern World (MW); and (vi) Science, Technology and Society(STS). Each student is requested to take two modules for written examination. Within these two modules, he/she has to choose one of them for school based assessment (SBA) by completing an individual project. The weightings of written exam and SBA are respectively 80% and 20%.

The aims of requiring students to prepare project reports are to:

- encourage and involve them in doing research on their own;
- give them credit for initiating tasks and assuming responsibility for organizing their own work;
- stimulate a sense of exploration and discovery;
- emphasize the learning process and not just the learning outcome; and
- foster teacher-student interaction in the learning process.

Individual student projects will be marked by school teachers. The marks and projects will be submitted to Hong Kong Examinations and Assessment Authority for moderation and further process.

The Reason for Moderation

The main reason for having moderation of SBA marks from schools is to ensure the fairness. Teachers know their students well and thus are best placed to judge their performance. In consultation with their colleagues, they can reliably judge the performance of all students within the school in a given subject. However, they are not necessarily aware of the standards of performance across all schools. Despite training in carrying out IES, and even given that teachers will assess students on the same tasks and using the same assessment criteria, teachers in one school may be harsher or more lenient in their judgments than teachers in other schools. They may also vary in the awarded mark ranges. To address these potential problems, the HKEAA (like most other examination authorities) makes use of various methods for ‘moderating’ assessments submitted by different schools, with an aim to ensuring the comparability of IES scores across schools.

Expert Judgment in the Moderation Method

The moderation method aims to achieve comparability across schools by adjusting

- the average of IES scores of students from a given school; and
- the spread of IES scores of students from the school

with reference to another measure, which could genuinely reflect student performance on SBA (see [2], [4], and [5]). Some common possible measures that could be used for moderation are written exam results of students or expert judgment on sampled student projects. The SBA in ASL LS requires each student to complete a study on a social issue. The whole process includes a number of steps, including the following.

- Develop project plan and project proposal
- Plan the enquiry
- Collect data
- Process and analyse data
- Draft report
- Finalise report

The process extends over a number of months, starting from Form 6 to Form 7 in secondary school. It can be seen that the nature of SBA is quite different from that of written exam. In this regard, expert judgment of sampled student projects is preferred to be used for moderation of SBA marks in ASL LS.

The whole school is formed as a moderation group. We require mark standardization process to be conducted within a school across different modules. Five sample projects covering different levels of performance are selected from each school using stratified sampling. The accuracy of sampling will be explored using simulation in the next section. After sampling projects from schools, each of these projects will be double marked by two markers. Besides, double marking and discrepancy marking will be implemented to ensure the marking quality.

The school average of external assessors' marks is employed to determine school performance level and spread. However, there may have some drawbacks in directly using external assessors' marks of a school to estimate the school average and standard deviation of project marks. It is because only five student projects are sampled for calculating these statistics of a school whose size could be quite large, up to 40-60. The reliability of these statistics may be cast in doubt. Therefore a hybrid approach employing both expert judgment and statistical modeling known as Bayesian hierarchical method is suggested in order to increase the reliability of the estimations for moderation.

Bayesian Hierarchical Modeling in the Moderation Method

With respect to the problem of small sample size, Bayesian hierarchical statistical modeling is suggested to be employed so as to share information across different schools in order to increase the reliability of the estimations of school mean and spread of project marks.

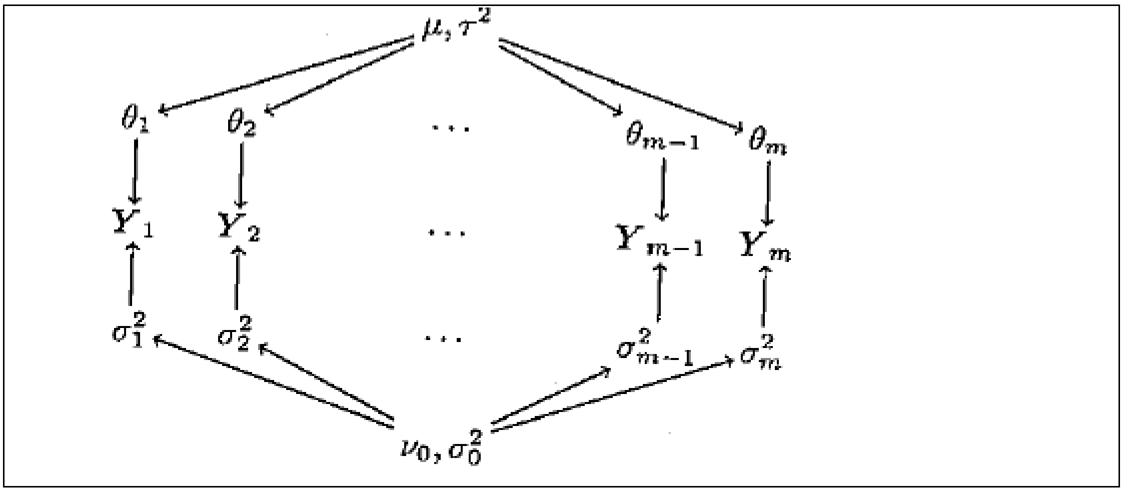
Let $\mathbf{Y_i}$ (a vector) be the marks from external assessors for a school i; i.e., $Y_{i,1}$, $Y_{i,2}$,$Y_{i,3}$,...,$Y_{i,ni}$. The number of students in the school is n_i. The Bayesian hierarchical model is set up as follows:

$Y_{i,1}$, $Y_{i,2}$,$Y_{i,3}$,...,$Y_{i,ni} \sim N(\theta_i,\sigma_i^2)$ for i = 1,…,m (i.e., there are m schools)

$\theta_i \sim N(\mu,\tau^2)$ for i = 1,…,m (i.e. all θ_i are sampled from a super-population)

$1/\sigma_i^2 \sim \text{gamma}(v_0/2, v_0\sigma_0^2/2)$ (i.e. all $\sigma_i^2$are sampled from a super-population)

The hierarchical structure could be represented in the diagram below (see [6]).



In Bayesian analysis, the parameters:μ, τ^2, v_0, and σ_0^2 are treated as random variables; instead of known/unknown constants. To conduct the Bayesian estimation, some non-informative priors $p(\mu)$, $p(\tau^2)$, $p(v_0)$, $p(\sigma_0^2)$ could be set up respectively for μ,τ^2, v_0, and σ_0^2. Based on such an approach, information could be shared across schools when estimating θ_i and σ_i^2. For schools with small sample sizes and/or extreme values, the estimates of θ_i and σ_i^2 will be pulled towards the corresponding overall estimates (μ and σ_0^2). Standard algorithms using Markov Chain Monte Carlo (MCMC) (see [1] and [3]) method are available for conducting Bayesian hierarchical modeling and some empirical results are presented below.

Simulation and Empirical Results

Simulation Results for Stratified Sampling

Basically, stratified sampling is employed for selecting student projects from a school. The steps are sketched below.

- Sort the school projects are in terms of projects marks in an increasing/ decreasing order.
- Sub-divide the list of sorted projects into 5 groups of more or less the same size.
- Within each group, randomly select a student project from it.

We use simulation to validate the precision of the sampling mechanism, For a 'simulated' school, whose project data are generated from a normal distribution with an appropriate mark range, the above sampling steps could be repeated for a thousand times. The simulated results from 1000 replications can be compared with the known values for the school mean and school standard deviation of project marks. The results are tabulated below. From the table, it can be observed that in general, the sampling results are quite close to the 'true' ones. The sample standard deviations are less accurate, as compared with the sample means. Moreover, the accuracy of the sample standard deviations is decreasing as the school size is getting larger. It is not a surprise as we only have 5 samples from a school. Nevertheless, the 'true' values of school standard deviations are still within the corresponding 95% C.I.

Table 1: Simulated results of stratified sampling using 1000 replications:

School Size	'True' School Mean	Average of Simulated School Means (95% C.I.)	'True' School Standard Deviation	Average of Simulated School Std. Dev. (95% C.I.)
7	36.20	36.15 (33.86, 38.30)	7.03	7.20 (6.13, 8.29)
9	34.83	34.76 (33.22, 36.55)	7.37	7.75 (6.41, 8.76)
12	37.46	37.48 (34.93, 39.36)	8.44	8.97 (6.70, 11.07)
17	38.57	38.60 (36.58, 40.51)	8.20	8.77 (6.66,10.65)
27	36.62	36.65 (34.35, 38.59)	9.08	9.79 (7.35, 12.29)
37	36.33	36.34 (34.26, 38.09)	8.96	9.72 (7.08, 12.14)
47	36.43	36.45 (34.48, 38.17)	8.81	9.61 (7.14, 12.05)
57	37.56	37.62 (35.49, 40.33)	9.22	9.96 (7.01, 14.05)

In addition to sampling errors, there could be non-sampling errors due to marking process conducted by external assessors. This kind of errors could be controlled by using Bayesian hierarchical modeling, of which empirical results are presented below.

Empirical Results of Bayesian Hierarchical Modeling

We select schools offering AS Liberal Studies in 2009 where a number of projects have been marked by external assessors. In the study, the number of the schools totally amounts to 98, in which there are 1208 candidates involved. The average ($\overline{y_i}$) and standard deviation ($sd_{e(i)}$) of each school are directly complied using external assessors' marks. On the other hand, Bayesian estimates for school average (θ_i) and school standard deviation (σ_i) could be obtained using the Bayesian hierarchical modeling with the data. The results for the schools are summarized below.

Table 2: Comparison of school averages and standard deviations with corresponding Bayesian estimates

Statistics	**School Average based on External Assessors' Marks** $\overline{y_i}$	**Bayesian Estimate of School Average** θ_i	**School Std. Dev. based on External Assessors' Marks** $sd_{e(i)}$	**Bayesian Estimates of School Standard Dev.** σ_i
Average	30.66	30.68	5.06	6.09
Std. Dev.	5.44	4.09	2.04	0.96
Minimum	18.50	22.54	0.82	4.46
1st Quartile	26.26	27.12	3.63	5.33
Median	31.55	31.26	4.96	5.96
3rd Quartile	34.23	33.37	6.42	6.64
Maximum	41.17	38.13	11.13	9.39

The shrinkage effect is mild for school average; while it is prominent for school standard deviation. The shrinkage effect is graphically demonstrated in the diagrams below.

Diagram 1: Distribution of school means of external assessors' marks and corresponding Bayesian estimates

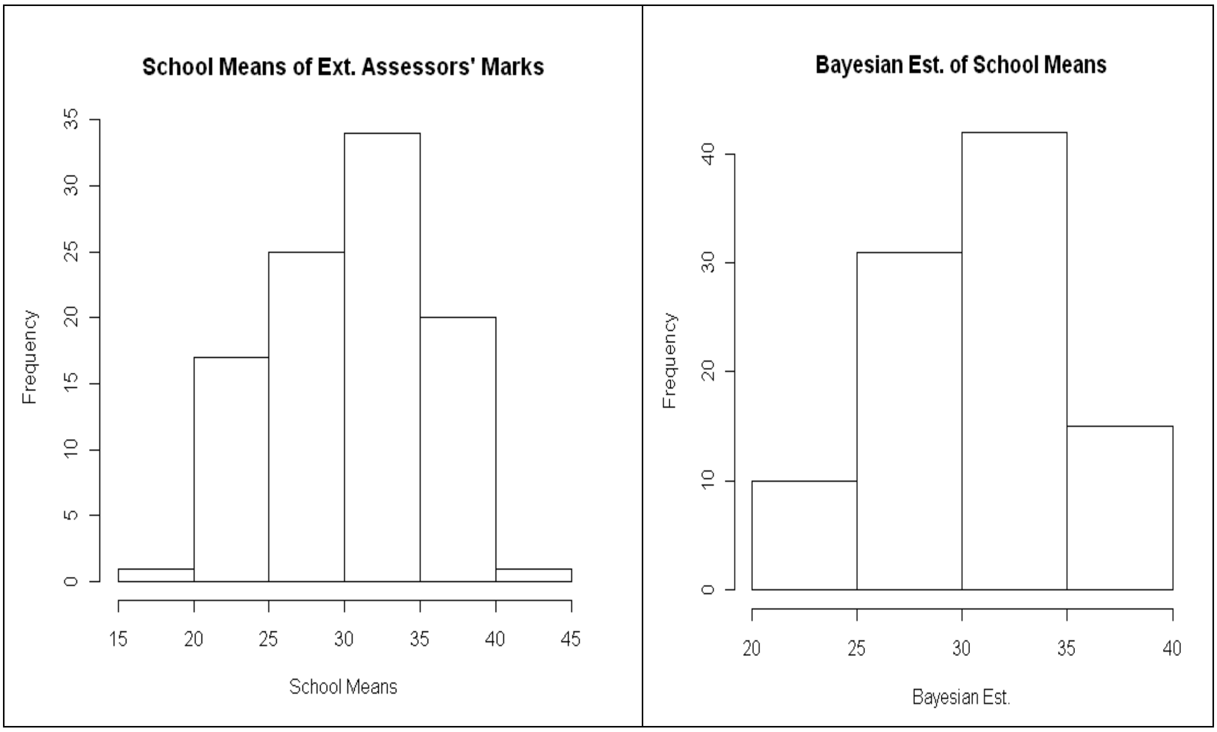
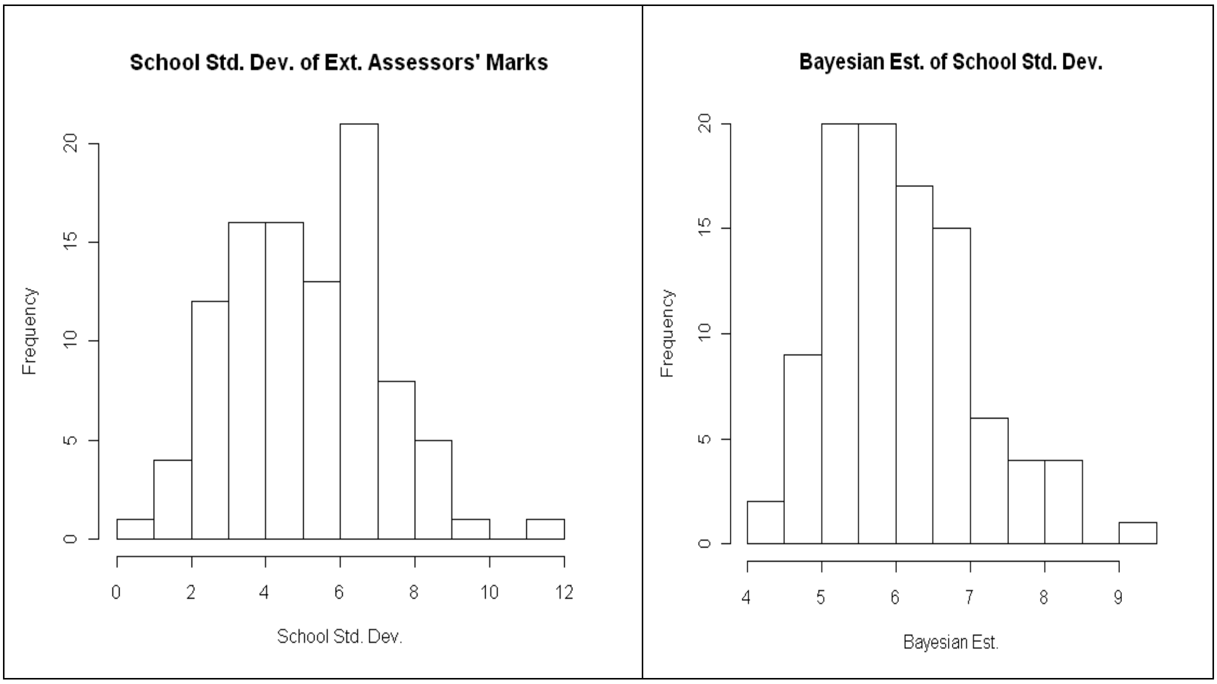


Diagram 2: Distribution of school standard deviations of external assessors' marks and corresponding Bayesian estimates



There are two main advantages of Bayesian hierarchical modeling, namely:

- This technique basically eliminates extreme values or outliers, and shrinks the values (of the average and standard deviation of external assessors' marks for a school) towards the overall estimates (μ and σ_0^2). The shrinkage magnitude for the result of a school depends on its sample size and its degree of extremeness.

- To counteract the problem of small sample size and high variability, sharing information across schools could stabilize the estimates for different schools. The 'effective' sample size is increased.

Discussion and Conclusion

In the paper, we mention the necessity of moderating SBA marks from schools in order to achieve comparability across schools. In view of SBA nature of ASL LS, expert judgment is preferred to be used as basis for moderation. However, due to time schedule and manpower resources, only limited number of samples could be obtained from a school for expert judgment. To improve the reliability of results, a hybrid approach supplementing expert judgment with statistical techniques is advocated. From empirical results, it can be observed that the employment of statistical modeling is a remedial measure for the problem of small sample size and safeguards the occurrence of extreme values.

We expect the proposed approach is especially suitable for subjects whose SBA performance of a school may not be highly related with that of written exam. The hybrid approach will be implemented in 2010 ASL LS exam. We will review the final outcomes and examine school feedback on the moderation method to explore any room for further improvement.

References

[1] Andrew Gelman, John B. Carlin, Hal S. Stern and Donald B. Rubin, *Bayesian Data Analysis*, Second Edition, Chapman & Hall; 2003.

[2] Board of Studies, *Statistical Moderation of VCA Course*, November 1999.

[3] Dani Gamerman and Hedibert F. Lopes, *Markov Chain Monte Carlo: Stochastic Simulation for Bayesian Inference*, Second Edition, Chapman and Hall/CRC; 2006.

[4] Government of Western Australia Curriculum Council, *Statistical Moderation*, 1999.

[5] Hong Kong Examinations and Assessment Authority, *Statistical Moderation of School-based Assessment Scores (For Schools)* April, 2007.

[6] Peter D. Hoff, *A First Course in Bayesian Statistical Methods*, First edition, Springer; 2009.