Hong Kong Teachers' Centre Journal, Volume 12

Marking and grading procedures for 2012 HKDSE Liberal Studies examination

FUNG Tze Ho & TONG Chong Sze

Hong Kong Examinations and Assessment Authority

Abstract

Liberal Studies (LS) is a new core subject for all candidates attending the 2012 Hong Kong Diploma of Secondary Education (HKDSE) Examination. Standards-referenced reporting (SRR) is adopted to report candidate performance, in terms of levels (from 1 to 5). Some LS teachers expressed doubts after the announcement of the grading results of the 2012 HKDSE LS subject. To address these concerns, this paper aims at reviewing the essence of marking and grading procedures for the 2012 HKDSE LS Examination. It is expected that the public could have more confidence in the attainment levels conferred by the Authority after having a clear and overall picture about the whole procedure.

Keywords

Hong Kong Diploma of Secondary Education Examination, Liberal Studies, standards-referenced reporting, marking and grading

1. Introduction

Liberal Studies (LS) is a new core subject for all candidates attending the 2012 Hong Kong Diploma of Secondary Education (HKDSE) Examination. In the HKDSE Examination, every subject adopts standards-referenced reporting (SRR) to report candidates' assessment results. In SRR, candidates' assessment results are reported, in terms of levels (from 1 to 5) with reference to explicit and fixed standards of performance stipulated as a set of level descriptors for a given subject. SRR has been adopted in Chinese Language and English Language of the Hong Kong Certificate of Education Examination (HKCEE) since 2007. Some LS teachers raised concerns after the announcement of the LS grading results in the 2012 HKDSE Examination. One of the controversial points is that

the percentage of LS candidates obtaining Level 2 or above amounts to some 90%, which may seem to be "unreasonably" high. In this regard, this paper aims at explaining the essence of marking and grading procedures of LS. It is expected that the public will have more confidence in the attainment levels conferred by the Hong Kong Examination and Assessment Authority after having knowledge about the marking and grading procedures.

In the following, the relevant marking and grading procedures of LS, and the related research studies and results will be highlighted. First, the LS assessment framework will be outlined. Secondly, marking arrangement for examination papers of LS will be mentioned; especially on the measures ensuring the reliability and validity in the marking process. Thirdly, moderation process of school raw marks on SBA will be studied, which aims at ensuring fairness and across-school comparability. After discussing marking process of exam papers and moderation process of SBA raw marks, the grading process, which is an essential part to determine the cut scores for various performance levels, will be examined.

2. Assessment framework of LS

There are two components in the assessment of LS, namely: (i) Public Examination, and (ii) School-based Assessment. In the component of Public Examination, there are two papers – Paper 1 and Paper 2. The Public Examination component amounts to 80% of the total (Paper 1: 50% and Paper 2: 30%), and the SBA component amounts to the rest, that is 20%.

Paper 1 consists of data-response questions, all of which have to be answered. Data-response questions aim to assess abilities such as identification, application and analysis of given data. The data define the scope and reflect the complexity or controversial nature of the issues involved; and such kind of questions also reflects the cross-modular nature of the curriculum.

Paper 2 consists of three extended-response questions. Candidates are required to answer one question only. Extended-response questions with data as stimulus information provide a wider context for candidates to demonstrate various high-order skills, such as drawing critically on relevant experience, creative thinking, and communicating in a systematic manner.

In addition to attending the public examination, each candidate of LS is required to complete an Independent Enquiry Study (IES) on a selected social issue, which is adopted as the mode of SBA in LS. The IES extends over a certain period of time and requires students to demonstrate various skills, such as data gathering, and analysis and presentation of findings. The IES is divided into three stages. The first is a preparatory

stage during which a candidate formulates the project title, specifies the objectives, considers suitable method(s), decides on the mode of presentation, plans for the enquiry and collects feedback from his/her classmates on the project plan. The second stage mainly involves data collection and organization. The third stage is the completion of the product which includes analysis and evaluation of data, conclusions on the results of the enquiry and a reflection on the enquiry process.

3. Marking arrangement

3.1 Marker training

The HKDSE LS examination consists of open-ended questions focusing on the enquiry of current social issues in accordance with the nature of the subject. There was a worry that the number of qualified markers may not be sufficient for the subject, as LS is a new core subject for all candidates of HKDSE Examination. Therefore, the Authority conducted three rounds of marker training sessions in the year 2010-2011. During the first round (from October to December 2010), a total of 10 sessions were completed, and 569 teachers participated. The second round comprising 9 sessions, was conducted from January to February 2011 and 538 teachers participated. The third round was between June and October 2011 and 594 teachers participated in 9 training sessions. Each training session comprised a 3-hour markers' meeting and post-meeting individual marking at the Assessment Centre. The training aimed to:

- allow teachers to experience the marking process, including the markers' meeting and the marking standardization process;
- provide opportunities for teachers to better understand the marking criteria and the standards of HKDSE LS;
- prepare teachers to be HKDSE LS markers and Assistant Examiners (AEs);
- familiarise teachers with the Onscreen Marking (OSM) system;
- collect marking statistics of teacher participants to facilitate the selection of markers for the live examination.

During the markers' meeting, participants were briefed of the marking criteria, standards and marking guidelines, illustrated by authentic performance in the sample scripts. Participants trial-marked some sample scripts. The scripts were then discussed in group meetings led by AEs who were experienced LS markers. Through the group discussions, with group size kept at 15 at most, participants aligned their marking standards and further clarified the marking criteria.

After the markers' meeting, participants marked 15 scripts of Papers 1 and 2 respectively on their own at the Assessment Centre. The marks of these scripts had been standardised by experienced markers in a previous exercise. Marking statistics, comparing

the characteristics of marks awarded by participants with that by experienced markers, were computed and sent back to participants as feedback. Marking statistics on the following aspects were discerned.

- *Mean of Mark Discrepancies*: This is the average of the discrepancies between the marks awarded by the participant and that of experienced markers.
- *Standard Deviation of Mark Discrepancies*: This is the variation of the discrepancies between the marks awarded by the participant and that of experienced markers reflecting the marking consistency; i.e., the lower the figure, the higher the consistency of marking performance.
- *Difference between the Mark Range of the Participant and that of Experienced Markers*: This shows the degree of discrimination relative to that of experienced markers.
- *Correlation between Marks of the Participant and that of Experienced Markers*: This indicates the degree of agreement between the marks awarded by the participant and those awarded by experienced markers, in terms of the rank order.

In the first round, amongst the 569 participants (681 enrolled) of the training, 383 joined the individual marking after the markers' meeting. In the second round, out of the 538 participants (606 enrolled), 394 joined the individual marking. During the last round of training, 542 out of the 594 participants (660 enrolled) completed the individual marking. Therefore, a total of 1,319 teachers participated in both the markers' meeting and individual marking. The following tables show the overall picture of marking statistics for the participants of individual marking in different rounds:

Table 1a: The averages of the statistical measures on marking performance of the participants of individual marking (Paper 1)

Statistical measure on marking performance (Max mark approx. = 20)	**1st round**	**2nd round**	**3rd round**	**Overall**
Mean of mark discrepancies	1.83	1.93	1.91	1.89
Standard deviation of mark discrepancies	2.36	2.39	2.40	2.39
Difference between the mark range of the participant and that of experienced markers	-0.37	-0.46	-0.42	-0.42
Correlation between marks of the participant and that of experienced markers	0.84	0.84	0.81	0.83

Table 1b: The percentages of the participants of individual marking fulfilling certain criteria (Paper 1)

Criterion	**1st round**	**2nd round**	**3rd round**	**Overall**
Mean of mark discrepancies between 4 marks and -4 marks[(a)]	97.65	96.95	97.35	97.32
Standard deviation of mark discrepancies less than 2 marks[(b)]	19.84	18.27	18.71	18.94
Difference between the mark ranges within ± 4 marks[(c)]	94.78	93.65	94.90	94.44
Correlation greater than or equal to 0.7[(d)]	97.13	96.45	96.03	96.54

Notes:

(a) 4 marks were determined as the thresholds for mean of mark discrepancies by considering the need of third marking, and corresponding resources available and time constraints.

(b) 2 marks were determined as the thresholds for standard deviation of mark discrepancies by considering the need of third marking, and corresponding resources available and time constraints.

(c) Provided that the variations of marks assigned are identical between two markers, it can be shown that the difference in mark range being greater than 4 is rare, with chance being equal to some 0.15.

(d) As a rule of thumb, in general correlation greater than or equal to 0.7 is regarded as high.

Table 2a: The averages of the statistical measures on marking performance of the participants of individual marking (Paper 2)

Statistical measure on marking performance (Max mark approx. = 20)	**1st round**	**2nd round**	**3rd round**	**Overall**
Mean of mark discrepancies	0.88	0.90	1.15	1.00
Standard deviation of mark discrepancies	2.48	2.47	2.41	2.45
Difference between the mark range of the participant and that of experienced markers	1.37	1.24	1.22	1.27
Correlation between marks of the participants and that of experienced markers	0.72	0.73	0.74	0.73

Table 2b: The percentages of the participants of individual marking fulfilling certain criteria (Paper 2)

Criterion	**1st round**	**2nd round**	**3rd round**	**Overall**
Mean of mark discrepancy between 4 marks and -4 marks	96.87	96.70	96.98	96.85
Standard deviation of mark discrepancy less than 2 marks	18.54	18.02	20.42	18.99
Difference between the mark ranges within ± 4 marks	89.56	88.32	88.66	88.85
Correlation greater than or equal to 0.7	65.27	69.29	72.21	68.92

For Paper 1, the averages in Table 1a displayed similar patterns in all the three rounds. The means of mark discrepancies were well within the "acceptable" level; i.e., below 4 marks. The mark ranges of the participants were just slightly smaller than that of the experienced markers, with an average for all participants being equal to -0.42 marks. The correlation was high with the overall figure being equal to 0.83. However, the averages of the standard deviations of mark discrepancies throughout these rounds were quite large; i.e., greater than 2 marks.

From Table 2a, it was observed that the performance of the participants in Paper 2 was quite similar to that in Paper 1, with good performance on average in terms of the mean of mark discrepancies, mark range and correlation, but slight under-performance for the item of standard deviation of mark discrepancies. The mean of mark discrepancies for Paper 2 was much closer to zero, though the correlation was lower than that in Paper 1.

From Tables 1b and 2b, an overwhelming majority performed satisfactorily in terms of the mean of mark discrepancies and mark range. The majority awarded marks that correlated well with that of the experienced markers, though the percentage of participants performing well in this aspect was much higher in Paper 1. For Paper 2, the percentage of participants with acceptable performance in terms of correlation and standard deviation of mark discrepancies had a slight increase from the first to the third Round. Based on these marking statistics, the percentage of discrepancy marking for Papers 1 and Paper 2 was (roughly) estimated to be around 20% for the live examination, which would be taken into consideration for manpower arrangement.

In a nutshell, a total of 28 training sessions were conducted in 2011-2012. A total of 1,319 teachers experienced the whole marking process and were familiarised with the OSM system. To facilitate the selection of markers, Principal Component Analysis (PCA) has been employed to derive an integrated marking performance indicator based on the four marking statistics so as to maximise the discrimination power. In addition

to the marking statistics, other factors, such as previous marking experiences, would be considered when selecting markers for the live examination of LS. A large majority of these participants performed satisfactorily with reference to a number of marking statistics; especially on the mean of mark discrepancies, difference between the mark ranges and correlation. This indicated that they were able to grasp the marking criteria and adopt the marking standard reasonably close to our experienced markers. On the other hand, there was room for improvement in the consistency of marking performance of these participants. In addition, 23 experienced LS teachers served as facilitators in group meetings and gained experience as AEs.

3.2 Markers' meeting and onscreen marking

Immediately after the completion of the LS public examination, the marking process was started. Markers' meetings with recruited markers were arranged in order to standardise the marking criteria and standards. Before the markers' meetings, a representative sample of candidate scripts was selected and marked by the Chief Examiner and a group experienced senior AEs whereby the consensus on marking standards and marking criteria were arrived at through professional discussion. Some of these standardised scripts were used for marking standardization, training and qualifying purposes. After the markers' meetings, markers then marked another set of standardised scripts which were used for testing whether they could grasp the marking standards and marking criteria properly so as to obtain the markers' qualification. Only those qualified markers would be allowed to mark scripts of the live examination.

In addition to manual procedures for ensuring marking quality, the Authority adopts innovative and advanced technologies to enhance the marking performance. In 2005, the Authority received funding from the government to modernise its information technology infrastructure, and to introduce OSM to improve the security, quality, reliability and efficiency of marking. The marking procedures with the use of OSM are outlined below:

- Examinations for candidates conducted;
- Examination scripts collected;
- Examination scripts scanned and images saved;
- Images of answers distributed to markers for viewing and marking via secure intranet system at designated Assessment Centres;
- Marks at question level and annotations by markers captured by the onscreen marking system.

For security reasons, marking is conducted at designated Assessment Centres. The primary function of these Assessment Centres is to facilitate onscreen marking of public examinations but they will also be used for the delivery and marking of a wide range of examinations and assessments, such as the Territory-wide System Assessment and a

variety of computer-based examinations. Moreover, facilities will be available for the training of examiners, markers, teachers and other assessment staff. The advantages of using OSM include the following aspects:

- Security: Secure storage of scanned images of scripts, and elimination of the physical movement of massive scripts;
- Marking: Real-time monitoring of marking consistency and quality control of marking, and flexible allocation of questions to markers;
- Efficiency: More efficient and flexible script management processes, and higher efficiency in mark calculation;
- Accuracy: Reduction of errors arising from mark entries, and elimination of errors associated with manual mark calculation and recording processes;
- Data Availability: More detailed analysis of candidates' performance, and more information on responses to individual questions and better feedback regarding candidates' performance.

In view of all the aspects mentioned above, OSM is considered as a better alternative to the conventional paper-based marking (PBM). Concerning marking quality, with the use of OSM a marker's performance could be continuously monitored by comparing his/her marks awarded on standardised scripts with that of experienced markers. Thus, marking problems identified could be rectified at an early stage. Besides, it also facilitates the sample checking conducted by AEs on certain scripts of each marker.

The Authority first introduced OSM in the 2007 HKCEE English writing paper. Afterwards, OSM was being implemented gradually in marking exam scripts for a number of subjects. To ensure that there is no adverse effect of OSM on the marking performance, the Authority has initiated a number of studies with tertiary institutes comparing OSM with PBM. A study (Coniam, 2009a, 2009b) examined English language essay scripts selected from the 2007 HKCEE English Language Paper 1B (Writing). To compare OSM with PBM, 30 markers, who had good rater statistics, were arranged to remark on paper 100 scripts, which they had marked onscreen nine months before. After the remarking, they were requested to complete a questionnaire in order to collect feedback on the exercise. From the questionnaire data, it was suggested that technologically, raters had no problems with OSM. Attitudinal differences surfaced, however, between new raters who had solely rated on screen as against experienced raters who had solely adopted PBM in their previous experiences. New raters felt that having to travel to a special marking centre was less of an inconvenience than did old raters. New raters, additionally, expressed a preference to mark on screen rather than on paper.

The statistical analysis of remarking data was conducted from two perspectives. The first involved classical measurement statistics. Correlations between the two forms of rating and the amount of discrepancy scripts (where a third rating was required) suggested

that no bias existed favouring either form of marking. Secondly, using multi-faceted Rasch measurement (MFRM), a five-faceted design was employed, modeling raters, test takers, input prompt materials, rating scales, and, especially, the marking medium. Results showed that all factors generally exhibited good data fit. For the method of marking - the major facet for investigation, the corresponding logit values of both methods were very close to zero. Therefore the hypothesis that the methods of marking (OSM and PBM) did not interfere scores awarded by markers was accepted.

There is another study (Coniam, 2010) which has similar objectives as the first one; but the subject concerned is Advanced Supplementary Level (ASL) Liberal Studies. The study involved 14 markers who had previously marked ASL Liberal Studies scripts on screen in the 2009 Hong Kong Advanced Level Examination. In the study, the 14 markers remarked on paper a number of the scripts that they had marked on screen in the 2009 examination. Using multi-faceted Rasch analysis, a five-faceted design was employed to model markers, test takers, input questions, rating scales and the marking medium. Results showed that all factors generally exhibited good data fit and suggested that the scores from OSM could be considered as reliable as those obtained from PBM.

3.3 Double marking arrangement

With regard to marking reliability, one of the public concerns is that there may be a considerable degree of variability when marking open-ended questions of LS. In this regard, the Authority has decided to adopt double marking in LS public examination. Any LS question of a candidate will be primarily marked by two markers. In case that prominent discrepancy occurs between the two markers' marks, third marking (i.e., discrepancy marking) will be undertaken. The average of the closest pair of marks[1] will be taken as the final mark of the question concerned. Fourth marking may be involved, if necessary, to settle down any controversies. Due to the use of OSM, which facilitates immediate distribution of scripts and flexible allocation of questions, double marking could be conducted on question basis. The four questions in Papers 1 and 2 attempted by a candidate in the public examination of LS will be marked by at least eight markers. Such an arrangement eliminates the chance that a candidate's assessment result will be dominated by a single marker who may be too harsh or too lenient.

The Authority had undertaken a study (HKEAA, 2011a) to examine the impact when adopting double marking in the LS questions. In the study, four data-response questions and four extended-response questions, and the corresponding marking guidelines were prepared in both Chinese and English. The full mark of each of these questions was more or less 20. These questions were attempted by some 1,300 students from 15 schools

1 In OSM, the sum of the closest pair of marks is compiled instead for the sake of computational convenience. This, in fact, implies that the full mark of a question is doubled.

covering a wide spectrum of performance levels. The student responses were marked by 18 markers using double marking (with discrepancy marking). Each student attempted one data-response question and one extended-response question, resulting totally 2,530 responses. For these 2,530 student responses, double marking was conducted. The corresponding statistics on marking discrepancies are shown below.

Table 3: Distribution of discrepancies in the study on double marking

Abs Diff	Count	Percent	Cumulative percent
0	413	16%	16%
1	749	30%	46%
2	592	23%	69%
3	368	15%	84%
4	226	9%	93%
5	100	4%	97%
6	51	2%	99%
7	20	1%	100%
8	9	0%	100%
9	1	0%	100%
10	1	0%	100%
ALL	2,530	100%	-

Some 16% of total responses, which had differences greater than three, required discrepancy marking. In general, third marking was already sufficient to ensure that the differences between the closest pairs of marks were less than or equal to three marks. There were only a small proportion of responses that required fourth marking. The corresponding distribution of discrepancies after discrepancy marking is tabulated below.

Table 4: Distribution of discrepancies after discrepancy marking in the study on double marking

Abs Diff	Count	Percent	Cumulative percent
0	510	20%	20%
1	911	36%	56%
2	724	29%	85%
3	385	15%	100%
ALL	2,530	100%	-

The closest pair of marks of a response was used for calculating the average, which was the final mark of the response. Provided that the "true" performance of a response did fall in between the closest pairs of marks, the difference between the final mark assigned and the "true" performance would be less than 1.5 marks; i.e., less than 10% of the full mark of the question concerned. The correlation between the marks in the closest pairs (retained after conducting double marking with discrepancy marking) was found to be equal to some 0.8. This reflected a high level of marking reliability.

In 2012 public examination of LS, it is found that the percentage of responses that requires discrepancy marking further decreases. It may be due to the fact that previous professional development courses and the targeted marker training sessions have familiarised school teachers with the marking criteria and standards of HKDSE LS.

4. SBA Moderation Process

4.1 The reasons of moderation

SBA is a salient feature of the HKDSE Examination. SBA refers to assessments administered in schools and marked by the students' own teachers. SBA in LS requires each student to carry out an Independent Enquiry Study (IES). The IES provides a valuable opportunity for students to independently carry out a focused enquiry into a contemporary issue of interest, and to present their views, ideas, findings, evaluations and personal reflections.

After receiving the raw SBA marks from schools, the Authority has to undertake the SBA moderation process. The main reason for carrying out moderation is to ensure the consistency of assessment standards across schools. Teachers know their students well and thus are best placed to judge their students' relative performance. However, they could not be aware of the standards of performance across all schools. Therefore, teachers in some schools may be harsher or more lenient in their judgment than teachers in other schools. Mark ranges of scores awarded in various schools may also be different from each other.

To resolve these problems, the Authority employs appropriate methods for "moderating" the raw SBA scores submitted by different schools to achieve the following:

- The comparability of SBA results across schools in order to ensure fairness for individual students and schools;
- The quality, reliability and validity of SBA results;
- Provision of useful feedback to schools for improving practice;
- In LS, the SBA moderation is conducted using statistical moderation based on examination results and supplemented with sample review.

4.2 Statistical moderation

Statistical moderation is particularly appropriate in situations where there is another measure available that can reflect SBA performance level. Typically this other measure will be students' performance in the public examination of that subject. An advantage of the method is that it can be carried out efficiently and impartially within a reasonable amount of time and resources. The key assumption is that the overall performance in the public examination of students in a school can properly reflect the SBA performance level of the same group of students. Generally speaking, this is a valid assumption in the context of many academic subjects in public examinations.

In the moderation process, the adjustments are applied only to school average and spread of raw SBA scores of students with reference to their public examination scores in the same school. Therefore, the ranking of students within a school remains unchanged after moderation. The school averages of examination scores are used to determine the corresponding performance levels on SBA, taking within-school correlations between students' raw SBA scores and examination scores into consideration (HKEAA, 2010).

4.3 Sample review

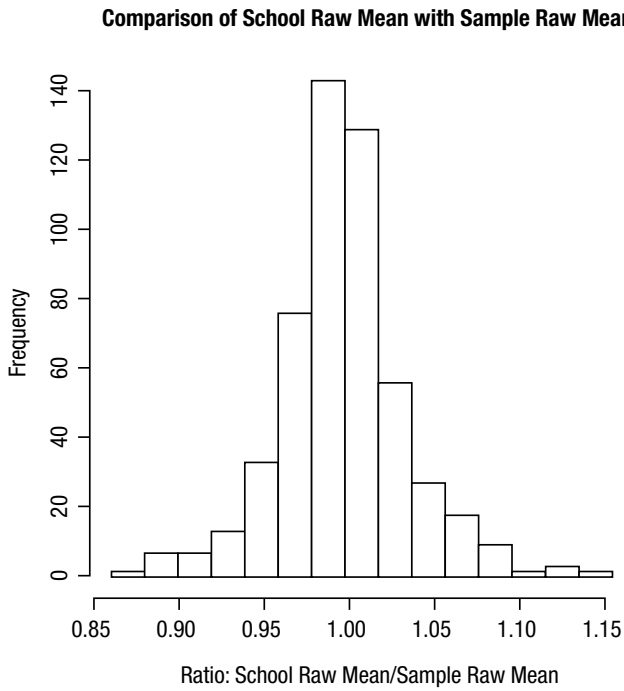
Some of the objectives of the SBA cannot be precisely assessed in the public examination due to different requirements. Moreover, students in SBA would possibly gain significant improvement under teachers' supervision due to the efficacy of assessment for learning. If only schools' public examination scores are used to adjust students' raw SBA scores, for some schools the statistical moderated results may not fully reflect the students' actual performance in the SBA; i.e., there may be some outlier schools whose statistically moderated scores differ greatly from the performance level demonstrated by students' SBA work. Therefore, for 2012 HKDSE LS, each school was required to submit six samples of students' work for reassessment which was conducted by a group of external assessors appointed by the Authority. The samples were chosen by the Authority using stratified random sampling. Students in each school were divided into a number of strata based on their raw SBA scores. Therefore, in each stratum the performance level of students on SBA should be similar with each other. Some students' work was then randomly chosen from each stratum. The stratified sampling method could ensure that a fairly small sample of students' work could adequately represent the full range of SBA performance of each school. For schools where only a few students were studying a particular subject, the work of all students had to be submitted.

All the LS samples were then reassessed with reference to the previous standardised exemplars and a set of stipulated assessment criteria. If prominent discrepancies between external assessors' scores and raw scores were observed, discrepancy marking would be conducted. It was observed that the discrepancy marking percentage was about 20% in 2012. The correlation between raw scores and results based on external assessment

amounted to 0.8. This reflected that the marking standards of school teachers were generally in line with that based on external assessment.

With regard to possible sampling variations, the ratio of school average of raw SBA scores to sample average of raw SBA scores was examined for each school. The distribution of these ratios of 523 schools is shown below.

Figure 1: Distribution of ratios of school means of raw SBA scores to sample means of raw SBA scores



The 5% percentile of the distribution was 0.94 and the 95% percentile was 1.06. It implied that sample raw means were very close to school raw means for most schools. In addition, it should be noted that the mean mark of sampled students' work from external assessment of a school would be adjusted upwards when sample raw mean was less than school raw mean; and vice versa. With such adjustments, it was expected that the sampling variations would be further reduced.

To further enhance the reliability of the estimations of means and spreads of SBA scores of schools based on external assessment, Bayesian hierarchical modeling was employed so as to share information across different schools. The model is briefly described below.

Let Y_i (a vector) be the marks based on external assessment of 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 ni. The Bayesian hierarchical model is set up as follows:

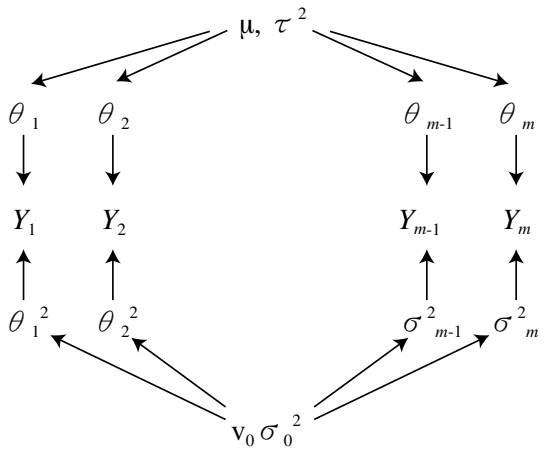
$Y_{i,1}, Y_{i,2}, Y_{i,3},...,Y_{i,ni} \sim \text{Normal}(\theta_i, \sigma_i^2)$
for i = 1,…,m (i.e., there are m schools)

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

$1/\sigma_i^2 \sim \text{Gamma}(v_0/2, v_0\sigma_0^2/2)$
for i = 1,…,m (i.e., all σ_i^2 are sampled from a super-population)

The model is graphically displayed in the figure below.

Figure 2: The structure of Bayesian hierarchical modeling showing relationship between data observed and parameters involved



In Bayesian analysis, the parameters: μ, τ^2, v_0, and σ_0^2 are treated as random variables. To conduct the Bayesian estimation, some non-informative priors p(μ), p(τ^2), p(v_0), p(σ_0^2) are set up respectively for μ, τ^2, v_0, and σ_0^2. Based on such a model, information could be shared across schools when estimating θ_i and σ_i^2. For schools with small sample sizes and/or extreme empirical values, the estimates of θ_i and σ_i^2 will be pulled towards the corresponding overall estimates (μ and σ_0^2). In general, algorithms using Markov Chain Monte Carlo (MCMC) method are deployed for estimation in Bayesian hierarchical modeling. It is well known that such a hierarchical model could reduce the estimation error (Berger, 1993; Hoff, 2010; Gelman et. al., 2003) in different applications. In addition, a simulation study has been undertaken to gauge the magnitude

of gain in accuracy when applying the model in the specific setting for SBA moderation (Fung, 2011). It is found that the total Mean Squared Error (MSE) in the estimation of school means could be reduced by some 30% using Bayesian hierarchical modeling, as compared with the one simply using sample means.

After consolidating the sample review result of a school, it was compared with the corresponding result from statistical moderation. Due to possible variations incurred in the sampling and remarking process, an appropriate tolerance limit was set when comparing the two results. If the difference was within the tolerance limit, the statistical moderation result would be adopted as the school performance level on SBA. If the difference exceeded the tolerance limit, appropriate adjustments would be made to the statistical moderation result with reference to the sample review result in order to determine the school performance level on SBA.

It is worth mentioning that in LS, the SBA marks of a student is divided into two parts, namely: (i) Task and (ii) Process. Only marks on the Task of a student will be moderated according to the procedures mentioned above. Marks on Process which includes students' effort in the IES will not be subject to moderation, as students' performance in this part may not be prominently associated with the examination results. Schools are expected to award the Process marks in accordance with the stipulated criteria. The Authority imposes quality control measures to ensure the fairness and reliability of the assessment on Process, which include monitoring by District Coordinators (DCs), providing feedback to schools and follow-up of any irregularities identified.

In 2012, it is observed that the mean of Process marks submitted by all schools is quite appropriate (i.e., not too high or too low) and the spread is reasonable. The moderated Task marks are then combined with the un-moderated Process marks to form the total SBA score for inclusion in the subject result. For the Task component, in 2012 53.3% of schools fall into the "within the expected range" category[2], while the marks of 21.5% of schools are higher than expected, and 25.1% lower than expected. Moreover, among the schools with marks higher or lower than expected, the majority only deviate slightly from the expected[3]. Thus, in 2012 the majority of schools falls into the "within the expected range" or "slightly higher/lower than expected" categories. It is supposed that teachers in these schools do have a good understanding about the marking standards.

---

2 Based on the difference between the means of the moderated and raw Task marks (D), a school is in the category of "within the expected range" when $0 \leq D < 3$ with full mark = 50.

3 The difference between the means of the moderated and raw Task marks is greater than or equal to 3 and less than 6 with full mark = 50.

5. Grading process based on professional expertise

Under SRR, a set of draft descriptors has been developed for each subject to describe how a candidate typically performs at a given level. The main purpose of grading is to determine the minimum score needed for a candidate to attain a given level. This minimum score is known as the cut score.

The HKDSE grading procedures include a series of tasks (HKEAA, 2011b) that begins before the actual marking of scripts. For any given subject, a panel of expert judges, which comprises the subject manager(s), the chief examiner(s) and selected assistant examiner(s) or markers from the individual components, is responsible for conducting the series of grading tasks, including: (i) sample script selection, (ii) marking standardization, (iii) post-marking exercise, and (iv) panel of judges grading meeting.

After the 2012 public examination of LS, some samples that could illustrate performance particularly well in relation to the level descriptors were selected. After script selection, the panel discussed the scores to be awarded to discrete points in the sample scripts. These marked scripts were used as standardisation scripts for marking.

After the completion of marking and moderation of SBA scores, the panel considered the selected written examination exemplars and SBA samples with reference to the level descriptors, and the previous released samples. The objective of the discussion was to make provisional grading recommendations (including preliminary cut score ranges) on each examination paper and SBA component through expert judgment based on samples of performance.

In the panel judges grading meeting, panel members re-considered the level descriptors, question requirements, marking guidelines and a number of representative samples as well as a range of recommended cut scores for each level. Panel judges exchanged their views led by the Chief Examiner. With a number of rounds of discussions, they finally agreed on preliminary cut scores for each paper and SBA component, and for the subject. In determining the cut scores, consideration was made to the actual performance of candidates in relation to

- the level descriptors;
- performance samples from the HKDSE SRR Information Packages (HKEAA, 2009);
- marked live scripts selected;
- feedback from markers on the level of difficulty of each particular examination paper;
- performance statistics of current papers and SBA component.

During this meeting, the panel of judges investigated the impact of amending the cut scores for each examination paper on subject grade distributions. Finally, the panel of judges decided on their recommendations for the cut scores for LS.

A senior management team led by the Secretary General of the Authority reviewed and decided on the cut scores based on the recommendations made by the panel of LS, and submitted the cut scores from the panel of LS to the Public Examinations Board (PEB) for further discussion and endorsement. In 2012, after discussion in PEB it was endorsed that the recommendations made by the panel of LS were strictly followed without any adjustments. The cut scores for Level 5** and Level 5* were set with reference to the percentage in mark distribution so that Level 5** was awarded to the highest-achieving 10% (approximately) of Level 5 candidates and Level 5* was awarded to the next highest-achieving 30% (approximately) of Level 5 candidates.

6. Conclusions

In this paper[4], it is highlighted that the Authority has taken stringent measures to ensure the quality of marking and grading procedures adopted in HKDSE Examination of LS. Relevant researches were conducted to examine the impacts of the new measures as far as possible. It is expected that after having an overall picture of the marking and grading procedures, the public will have more confidence in the attainment levels conferred by the Authority.

Currently, the Authority is now collecting opinions and feedback from various stakeholders on the assessment framework of LS in order to strive for further improvement in the future.

References

Berger, J. O. (1993). *Statistical decision theory and Bayesian analysis* (2nd ed.). Springer.

Coniam, D. (2009a). Validating onscreen marking in Hong Kong. *Asia Pacific Education Review, 11*(3), 423-431.

Coniam, D. (2009b). Examining negative attitudes towards onscreen marking in Hong Kong. *CUHK Education Journal, 37*(1-2), 71-87.

4 To facilitate the access to the content by the public, the paper is also available from the website of the Authority.

Coniam, D. (2010). Markers' perceptions regarding the onscreen marking of Liberal Studies in the Hong Kong public examination system. *Asia Pacific Journal of Education, 30*(3), 249-271.

Fung, T. H. (2011, July). *Simulation study on the use of hierarchical Bayesian modeling in expert judgment for SBA Moderation*. Paper presented at the 76th Annual and the 17th International Meeting of the Psychometric Society, Hong Kong.

Gelman, A., Carlin, J. B., Stern, H. S., & Rubin, D. B. (2003). *Bayesian data analysis* (2nd ed.). London: Chapman & Hall/CRC.

Hoff, P. D. (2010). *A first course in Bayesian statistical methods* (1st ed.). Springer.

Hong Kong Examinations and Assessment Authority (HKEAA). (2009). *Liberal Studies: Standards-referenced reporting information package*. Further information available from: http://www.hkeaa.edu.hk/en/resources/publications/list_of_publications/hkdse_srr_pub/

Hong Kong Examinations and Assessment Authority (HKEAA). (2010). *Moderation of school-based assessment scores in the HKDSE booklet*. Retrieved from: http://www.hkeaa.edu.hk/en/Resources/leaflets/

Hong Kong Examinations and Assessment Authority (HKEAA). (2011a). *Study in double marking of Hong Kong Diploma of Secondary Education Liberal Studies practice questions' answer script*s. Retrieved from: http://www.hkeaa.edu.hk/en/Resources/research/

Hong Kong Examinations and Assessment Authority (HKEAA). (2011b). *Grading procedures and standards-referenced reporting in the HKDSE examination*. Retrieved from: http://www.hkeaa.edu.hk/en/Resources/leaflets/

2012 年香港中學文憑通識教育科考試的閱卷與評級程序

馮子豪、唐創時
香港考試及評核局

摘要

2012 年香港中學文憑考試的考生必須修讀通識教育科。香港中學文憑考試採用水平參照模式匯報考生的表現，將考生表現分為各等級（1 至 5）。部份通識科教師對通識科考試評級結果表示疑慮。有見及此，本文回顧通識科考試的閱卷與評級程序，期望當大眾認識相關的程序後，將對考評局所發的資歷更具信心。

關鍵字

香港中學文憑考試，通識教育科，水平參照模式匯報，閱卷與評級