第２回 AP 勉強会

日時：平成２７年１２月１５日（火）　１０：００～１２：００
場所：岡山大学津島キャンパス　大学会館　アドミッションセンター会議室
出席者：（学外者）International Baccalaureate Office，Chief Assessment Officer
（学内者）田原ｾﾝﾀｰ長，飯塚教授，佐竹特任教授，Mahmood 准教授，Fast 講師

Diploma Program assessment—aims and approaches

Diploma Program (DP) assessment is a summative assessment, designed to record student achievement at, or towards the end of the DP course. However, many of the assessment instruments can also be used formatively during the course of teaching, particularly for internal assessment tasks.

1. Internal assessment and other non-examination components

Internal assessment can be an oral presentation or a discussion of research work and investigations. The assessment task reflects the purpose of the internal assessment, and emphasizes the skills involved. Certain features are common to all internal assessments. Firstly, internal assessment is a part of normal classroom teaching, which focuses on skills, not subject content. Activities used for internal assessment can be used to develop skills, and also contributes to the final assessment outcome. The teacher decides how to use an activity as part of the final summative assessment. Secondly, the teacher provides a certain level of support to the student for activities that contribute to the final assessment. When the task is a written piece of work, teachers generally discuss the topic with the student and gives advice on the first draft. Subsequent editing is done by the student, and the final work is submitted by the student. Thirdly, internal assessment is conducted by applying a fixed set of assessment criteria for each course. These criteria describe the kinds and levels of skills that must be addressed in the internal assessment.

2. Marking

The main aim of the IBO assessment process is to provide almost the same mark to a piece of work, regardless of which examiner marked it. Assessment is done in three main steps. First by appointing examiners who can mark consistently and objectively. Second, by checking the markings of all examiners in every examination session except the senior examiner. This is called “moderation”. The third method is by providing instructions to examiners through prior training about the administrative procedures to be

followed and how to allocate marks. The IBO uses two principal methods of mark allocation: Analytic Markschemes and Assessment Criteria (markbands).

2.1 Analytic Markschemes

Analytic markschemes are prepared for questions that expect a particular kind of response and/or a given final answer. These markschemes give specific instructions to examiners on how to break down the total mark for a question for different parts of the response, as suggested by the senior examining team. Markschemes also provide examiners with information on how to mark consistently, and about the different approaches that candidates might adopt and the common errors they might make. Some questions require the examiner to use their professional judgment in allocating marks to unexpected responses or alternative valid answers, but markschemes provide guidance on how to make this judgment.

2.2 Assessment criteria

The assessment criteria is applied when an assessment task is so open-ended that the variety of valid responses is too great to permit analytical markschemes. The assessment criteria does not refer to the specific content of a candidate's answer, but concentrates more on the generic skills that candidates are expected to demonstrate, regardless of the specific response. Both internal assessments and externally assessed non-examination tasks are marked using assessment criteria. In all cases where assessment criteria are applied, different achievements are awarded different marks. The total possible mark for a piece of work is arrived at by adding the maximum achievement level for each criterion. Greater importance is given to criteria considered more important by giving a greater number of marking. The approach used in DP assessment in the application of criterion achievement levels is a "best fit" model. The examiner chooses the achievement level that best matches the piece of work being marked. The highest level of any given criterion does not represent perfection. A number of examination tasks are marked according to the same assessment criteria for each examination session. Although the general nature of the task remains the same in each examination session, the specific requirements of each question may have implications for the way in which the assessment criteria should be applied. When assessment criteria are used with internal assessment, both teachers and moderators should refer to the reference is made to the published teacher

support materials, which contain examples of how to apply the criteria.

3. Standardization

In addition to markschemes and assessment criteria, assistant examiners also receive advice from senior examiners, by telephone and/or e-mail during the marking period itself. To address problems and reduce global bias arising from educational cultures and teaching styles around the world, senior examiners meet and review the scripts of a selection of candidates. This is called a standardization meeting. The purpose of this meeting is to make a small number of final additions and amendments to the markscheme and ensure that senior examiners have agreed to a certain interpretation of how the markscheme should be applied. The final decision is then passed on to all assistant examiners.

4. Markbands

When it is not possible to recognize separate assessment criteria, or when the work being assessed is so variable that a set of readily applicable criteria cannot be derived, a different approach is adopted called “Markbands”. These are used instead of separate criteria, which represent a single criterion applied to a certain piece of work, and judged as a whole. Each markband level corresponds to a number of marks. For example, one markband level may cover the range 6 to 10 marks. The examiner gives a mark from that range based on how well the work fits the relevant level within the markband scale. Based on research, there is little difference between the reliability of marking through markbands or assessment criteria.

5. Moderation of External Assessment

Moderation is a process of ranking. The purpose of moderation is to ensure that candidate marks, on the whole, are adjusted to more appropriate levels. Moderation is the principal tool for ensuring marking reliability and is conducted by a team of examiners. The principle examiner (PE) for a subject is often the chief examiner or deputy chief examiner or a former team leader (TL). Generally, a PE may also be the author of the examination paper or was greatly involved in setting that paper. A TL is an examiner who has past experience in marking consistently and accurately. For each subject, there is also an assistant examiner (AE). Each TL oversees up to 10 (AE). Every AE is allocated a minimum of 10 and a maximum of 20 scripts. After marking, the AE sends a sample of their marking to the TL, and not the PE, This sample is re-marked by the TL and a statistical comparison of the paired set

of marks determines whether the original examiner's marking is acceptable, sometimes with some slight adjustment, or maybe unacceptable completely.

5.1 Correlation criterion

The pairs of marks for each script undergo statistical analysis. One such statistical measure is the correlation coefficient which measures the consistency of the relationship between the two examiners' marking. A correlation coefficient of 0 indicates no relationship at all; a score of 1 indicates perfect consistency between the two examiners' marking. A coefficient of –1 indicates consistently opposing views between the two examiners. For an examiner's marking to be acceptable, the correlation coefficient must be at least 0.90, indicating a high level of agreement between the assistant examiner and team leader. If the correlation coefficient is less than 0.90, the AE's scripts are re-marked by a more reliable examiner.

5.2 Linear regression

A further analysis is carried out on each moderation sample using linear regression, which calculates the best-fitting straight line through the set of data points awarded by both the AE and the TL. On average, a regression line is calculated from the sample data by converting each mark (x) awarded by the AE into an equivalent mark (y) that the TL would, have given to that same data. For example, if the TL gives a mark of 46 to a script that the AE has given 42, the regression line, reflecting the average trend of marking difference, would moderate each mark of 42 into 43. For satisfactory moderation, the slope of the regression line must be between 0.5 and 1.5. If the slope of the line is too low, it means that the AE has spread the marks out too much, giving too few marks to weak work and too many marks to good work and the TL had to compress the AE's mark range considerably. If the slope is greater than 1.5, the line is too steep and means the AE has not differentiated sufficiently between poor and good candidate work and the TL had to expand the mark range.

5.3 Tailing

For marks in the top 20% and bottom 20% of the available mark range, the calculated regression line is modified and substituted by new "tailed" lines that link from the regression line to the maximum and minimum coordinates. This is done so that no mater how generous or strict an AE is, he/she cannot award marks below the minimum or above the maximum.

5.4 Other criteria

The difference between the mean AE sample mark and the mean TL sample mark must be less than 10% of the total mark available for that component. This means, if the total mark for a component is 30, the mean AE mark must be within three marks of the mean TL mark for the given sample.

6. Moderation of internal assessment

The moderation of internal assessment, where the original marking is done by classroom teachers, has a slightly different approach, indicated by the different titles given to the examiners—that is, principal moderator, senior moderator and assistant moderator. All internally assessed scripts are marked by applying assessment criteria. Moderators for most internal assessment components, except for language orals, are asked to judge whether the teacher's marking seems appropriate, rather than re-mark the marks awarded by the teacher. Teachers' marks are altered only when the moderator is sure they are inappropriate.

7. Grade awarding and aggregation

The grade award meeting (GAM) is the final stage of the assessment process for each component, which takes place about 35 days after the date of the examination. The team reviews the assessment components for the session, sets the grade boundaries for each of the higher level and standard level courses, resolves any outstanding issues relating to examiner marking, and carries out "at risking". The first task of the GAM is to reflect on the operation of each component. Senior examiners review the comments formally submitted by teachers about the examination papers and the reports from AE on the general nature of candidate responses. Following this, the team takes into consideration each component for which new boundaries must be set every session. The boundaries for internally assessed components, and externally marked non-examination components, are not revised each session. They are normally set only once, but new boundaries are set for each examination paper at each session. The change in boundary marks is normally slight because every effort is made to construct each new version of an examination paper at the same level of its predecessor.

7.1 Setting grade boundaries

The setting of grade boundaries (GB) requires long and careful consideration of information from the experienced judgment of senior examiners, statistical comparisons and the expectations of experienced teachers, who are familiar with the standards and know the candidates better than anyone.

The most significant GB for each examination paper are between grades 3 and 4, between grades 6 and 7, and between grades 2 and 3, judged and determined in that order. These GB have the greatest impact on candidates' progression into higher education. The principal means of setting judgmentally determined GB is by reviewing the standard of work expected of typical candidates at each grade.

7.2 Aggregation

When GB have been established for all components, to combine the marks from different components and form a percentage total mark, they need to be "scaled". Scaling is done to preserve the desired adjustment for each component. For example, a higher level course in a subject may be made up of 3 components and the model requires that component 1 contributes 50% to the final result, component 2, 30% and component 3, 20%. If component 2 is designed to have a total available mark of 90, then these marks, after moderation, would have to be scaled by dividing by three to achieve the required adjustment of 30%.

7.3 Grade distribution

After the aggregation of component marks and GB by computer processing, the GAM reviews the provisional subject grade distribution before confirming the final GB. Comparisons are made with previous years' results. A significant shift in subject grade distribution compared to the previous year requires explanation.

7.4 "At Risking"

When the final results are considered fair and correct, the senior examining team and other experienced examiners resolve outstanding issues relating to marking reliability. There may be a few examiners whose work needs re-marking. The main area of re-marking, however, will concentrate on "at risk" candidates, whose final grade is 2 or more grades worse than predicted and who are within 2 % of getting a better subject grade.

A second, much smaller, category of "at risk" candidates are those candidates who are only 1 grade below prediction, and within 2 marks of achieving that predicted grade. Ideally, all candidates within 2 marks of subject grade boundaries are reviewed to receive the correct grade. However, attention is focused on those categories the candidate is most likely to have suffered disadvantage from inaccuracy in marking and moderation.

8. The final award committee

The final award committee meets after all the grade award meetings have been held and just before the results are issued in early January/early July. This committee formally awards diplomas and certificates to those candidates who have met the requirements. It also authorizes appropriate action special cases.

9. Publication of results

Diploma and certificate results are published to schools and university admission systems on 5 January and 5 July each year for the two examination sessions. The results are sent electronically.

10. DP Scores and Grading

The IBDP assessment has internal and external components. Students are graded for the internal components from 7 (highest) to 1 (lowest) for each subject assessed throughout the course. Grade 1-7 reflects (poor, little, basic, good with some gap, sound, very good and excellent), respectively. The maximum possible total diploma score is 45 (6 courses x 7 points) in addition to 3 points for successful completion of the external components namely, Theory of Knowledge (TOK) and Extended Essay (EE) through written examinations at the end of the DP course. The other main core element Creative Action Service (CAS), although compulsory, does not contribute to the total point score. Students who gain at least 24 points are awarded the IBDP. About 80% of students receive the DP with an average score of 30 points. Although Higher level (HL) and Standard level (SL) courses offered in IB differ in scope, the IB philosophy is to assess both HL and SL against the same grade descriptor and are awarded the same number of points. A bilingual DP is awarded to either students who receive a grade of 3 or higher in 2 languages from language and literature studies or to students who receive a grade of 3 or higher in studies in language of literature and a grade of 3 or higher in an individual social or science subject in another language.