

IB assessment principles and practices: A guide to assessment for chief examiners and principal examiners

Introduction and overview

This guide is intended to provide an overview of the assessment processes that are relevant to chief and principal examiners. These roles are typically involved in all stages of the examination process and, while most of the following information will be familiar to you, we hope that by outlining the responsibilities of each role we will provide an understanding of the importance of each of the stages of assessment.

Are you joining up your various responsibilities?

Summative assessments need to provide meaningful, fair and reliable outcomes for candidates:

- **meaningful** because assessments should measure what was intended to be measured, as set out in the subject guide
- **fair** because the assessments should not favour particular groups of candidates
- **reliable** because a candidate should receive the same outcome regardless of who marked their assessment or on what day.

An assessment that is not meaningful, is not fair or results in unreliable outcomes is not fit for use as a high-stakes summative assessment.

A chief examiner's role is to oversee and to be responsible for the assessment of a subject, always being mindful of the purpose of the assessment. This role includes ensuring that standards are maintained from one year to the next.

A principal examiner is responsible for the marking standard of a component within a subject. All examiners marking a component are expected to mark in line with the principal examiner's standard because this ensures that all candidates receive the right outcome, that is, the outcome for a candidate is not dependent on which examiner marked their work.

The following diagram demonstrates the stages of the assessment cycle, which is a continuous circular flow and as such is intended to benefit from continuous improvement. This means continuously striving to make assessments more meaningful, more reliable and fairer.

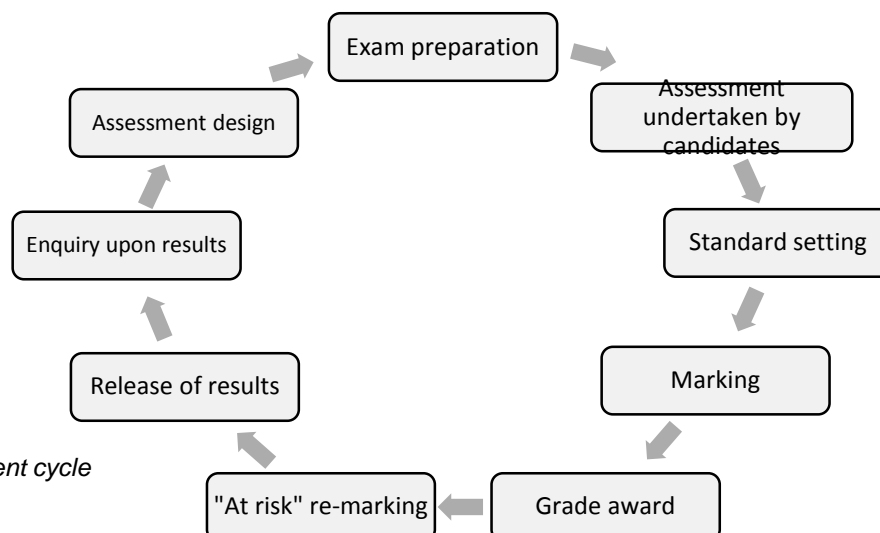


Figure 1. Assessment cycle

What is important in preparing the assessments and markschemes?

The chief and principal examiners, with support from other experienced examiners, are responsible for preparing examinations for each session, which may begin 18 months to 2 years before the examinations are taken. Thought should be given to all of the examinations and assessments, which make up the complete set required for the subject, to ensure they are balanced and together cover the full range of objectives. The senior examiner may write the whole examination or may compile the examination from questions submitted by other examiners.

Predictability is a challenging issue in creating assessments. Each examination should require candidates to tackle questions and/or tasks that are in some way different from what they have done before (to avoid negative backwash effects from teaching to the test), while also being comparable to previous examples (to allow candidates to prepare effectively). The IB seeks to test higher-order cognitive skills, so there must be questions that require candidates to solve a problem or think creatively to apply what they know in a new context, rather than just to proceed with well-rehearsed skills or to restate knowledge.

Questions and markschemes should be produced together to ensure that they are aligned in what they are trying to assess. The same effort should be put into the preparation of a markscheme as the questions during the exam-editing process. Even for those exams containing open-ended tasks/questions that are marked according to the same assessment criteria in each session, explanatory marking notes should be prepared that give guidance to examiners on how to apply the criteria in the context of each question. The markschemes are just as important to the integrity of the assessment process as the examinations. Each markscheme is much more than just a set of model answers; it provides guidance on how to mark common alternative approaches that candidates might adopt in answering a question and how to deal with commonly occurring errors or misconceptions that candidates might show.

The markscheme/marketing notes should be clear and should minimize the possibility of examiners approaching the marking differently. It is the main tool in communicating to examiners what is expected of them.

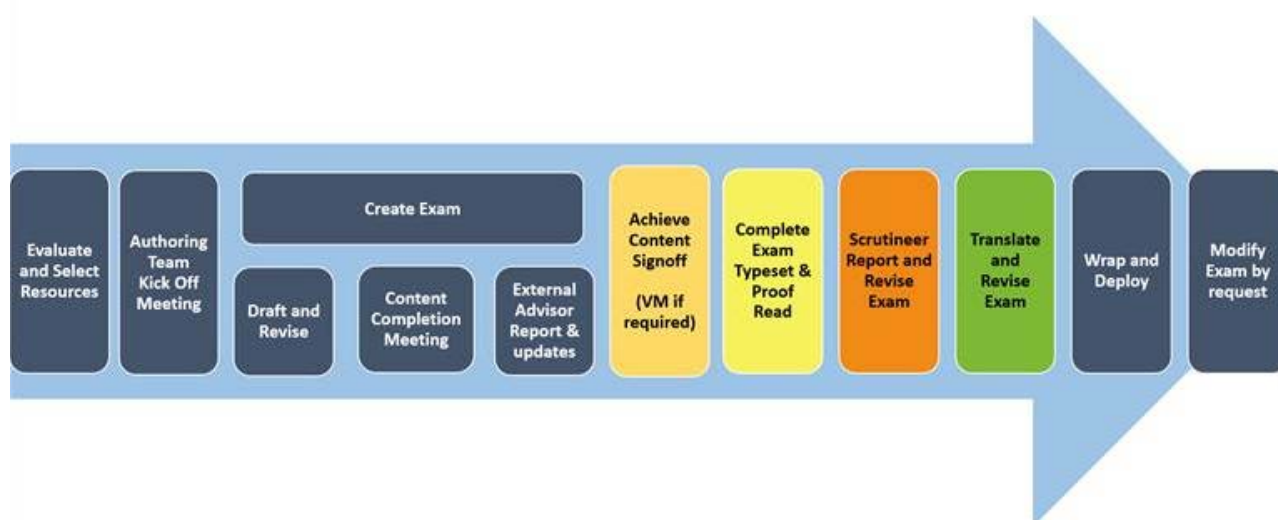


Figure 2. Preparing examinations.

How do you communicate your standard to your examiners?

Standardization is not a single meeting but a period of time that has several aims:

- for the principal examiner to set the standard for the assessment
- to test and refine the markscheme

- to produce definitively marked scripts (practice, qualification, seeds)
- to share understanding with all examiners
- to confirm examiners' understanding of the standard.

It is important to understand the different purposes of the three types of definitively marked scripts. The purpose of practice scripts is to support examiners in learning the marking standard, while qualification scripts are intended to demonstrate/prove examiners can mark to the correct standard. Seed scripts are used to demonstrate/prove examiners are continuing to mark to the correct standard.

The scripts that are selected should be of a clear standard as their purpose is to check whether examiners are applying the markscheme/criteria appropriately and marking to the principal examiner's standard, not to "catch them out". It is important that the scripts selected contain a range of responses and a range of marks in order to check that examiners have a wide and deep understanding of the markscheme/criteria, which is essential in order to recreate the principal examiner's marking standard.

The standard-setting stage has been successful if examiners have a good understanding of the standard and the IB has a set of definitively marked scripts (practice, qualification, seeds) to support examiners in learning and demonstrating they have mastered it.

How do you know your examiners understand your standard?

Marking reliability, the ability of an assessment process to provide almost the same mark to a piece of work regardless of which examiner marked it and on which occasion it was marked, is vitally important in ensuring fair outcomes for all candidates. In each component, the principal examiner sets the standard of marking and must communicate this standard clearly so that every examiner can mark in line with it. Deviating from this standard will result in unreliable outcomes for the candidates, which means the assessment will not achieve its purpose.

For e-marked components, the method used to ensure quality of marking is *seeding*. In the seeding process, examiners mark scripts that have been pre-marked by the principal examiner and the senior examining team so that the two sets of marks can be compared. If the differences between the two sets of marks are within a pre-determined "tolerance", the marking is considered to be acceptable. A tolerance is applied because it is recognized that there can be legitimate differences in the marks awarded by different examiners to the same piece of work, but it is important that these differences are minimized as much as possible so that there is minimal impact on candidate outcomes.

Tolerances will vary for different tasks to accommodate the fact that it is easier to agree on objective marks than subjective marks. In many cases it is reasonable to expect all examiners to give exactly the same mark, for example, when the answer is clearly correct or not. In longer, essay-type responses examiners should agree on the broad quality of the answer, but may disagree by a mark or two.

To ensure candidate work is marked accurately and to minimize the need for re-marking at the end of the marking period, examiners must mark a number of qualification scripts to prove that they are able to mark in line with the required standard before being able to access "live" scripts. Once they are marking live scripts, examiners are monitored throughout to check they are still marking to the required standard. Seeds are randomly introduced into examiners' marking for this purpose. Examiners must be unaware of which scripts are seeds so that they will mark them in the same way as any other script.

In every case the examiners' marks are compared to those awarded by the principal examiner in order to monitor whether candidates are receiving reliable outcomes. The IB does not take a view on what the correct mark should be.

Providing feedback to examiners

If an examiner's marks differ significantly from the definitive/agreed marks during live marking, it is important that the examiner is given immediate feedback on how and why marks have been awarded. The examiner is provided with a copy of the seed showing the definitive marks and the senior team's detailed annotations. If, despite this feedback, the examiner continues to mark outside the tolerance levels, he or she will be stopped temporarily from awarding marks until a senior examiner or team leader makes contact to discuss the marking.

The purpose of this discussion is for team leaders to provide quality feedback so that examiners can understand how to correct their marking. Seeds should be considered as opportunities for professional development for examiners. Most examiners will require feedback from seeds at some stage in their marking, and the feedback should be used to help ensure examiners meet the marking standard set by the principal examiner.

When the team leader is satisfied that the examiner is ready to start marking again, he or she will allow the examiner to access more scripts. However, if the examiner continues to mark seeds inaccurately, he or she will again be stopped from marking. This time, the information will be passed to IB staff and a decision will be made on whether the examiner should be permanently stopped from marking to prevent candidates from being awarded incorrect marks.

The proportion of examiners who mark seeds within the tolerance gives an indication of the level of marking reliability of a component. If a relatively small proportion of examiners is able to mark seeds within tolerance, the chief examiner and principal examiner should determine the cause of the low level of marking reliability and address it for future sessions. For example, if the markscheme was not clear enough in explaining how marks should be awarded, future markschemes should be reviewed to provide clearer guidance. If the assessment task itself does not enable reliable marking, this should inform the curriculum review so that the task can be changed or replaced entirely.

A great deal of care must be taken when setting the definitive marks for seeded scripts, not only because of their use in the quality assurance process but also because examiners are expected to use the marks and annotations to refine their marking. The potential impact of an error in a definitively marked script includes: examiners being unfairly stopped from marking, examiners not being stopped from marking (resulting in unfair outcomes for candidates), incorrect guidance being provided to examiners, and the incorrect mark for the candidates whose work has been definitively marked.

Question item groups

Some examinations will be marked by question item group (QIG), which may be part of a candidate's response or a whole response. If the examination is broken down into several parts, different examiners mark different sections. By dividing the examination into parts, each a separate QIG, examiners may choose which QIG to mark based on their knowledge and preference. Examiners' marking of QIGs is monitored using the seeding process.

Moderation of internal assessment

For internally assessed work, moderation is the principal tool for ensuring teachers' marking reliability. Unlike marking, moderation is intended to check that the teachers are applying the markscheme correctly, but assumes that teachers are consistent in their judgment.

How do you cope with contradictory evidence when awarding grades?

The grade award meeting represents the culmination of the assessment process from the candidates' point of view. Grade awarding gives meaning to the marks awarded to candidates; marks may vary between sessions as a result of the different questions, but the grade standard from previous sessions must be carried forward. For example, a grade 4 awarded in the current session represents the same standard as a grade 4 awarded in the last session and the one before that. To aid this consistency, boundary scripts from previous sessions are reviewed by the grade award team.

The first task of the grade award meeting is to reflect on the operation of each component. All participants will have been actively involved in marking at least one component, and most will have contributed to writing the examination papers. In addition to their own experience, senior examiners should review the comments formally submitted by teachers about the examination papers, reports from examiners about their perceptions of the papers and the statistical data from marking the papers. This information provides important background to help agree appropriate grade boundaries.

Next, the team should consider each component for which new boundaries must be set every session. The change in boundary marks should normally be slight because every effort should be made to construct each new version of an examination about the same level of overall difficulty as its predecessor. It is the chief examiner's responsibility to take into account all the available evidence (feedback on the performance of the examination and the statistical information as well as the senior team's judgments) to decide on the recommended boundary mark that will best carry forward the standard.

The boundaries for internally assessed components and externally marked non-examination components are not revised each session, since the requirements of the task are the same unless the course has changed. In these cases, the boundaries are normally set only once at the start of the course, although they can be reviewed each session.

When the final results are generally deemed fair and correct, the senior examining team and other experienced examiners can resolve outstanding issues relating to marking reliability. The main area of re-marking will concentrate on “at risk” candidates. Generally, these are candidates whose work was marked by examiners whose marking was flagged as potentially problematic, or candidates whose final grade is two or more grades below prediction and who are within two percentage marks of getting a better subject grade. Given that there is an error of measurement in marking, the accuracy of marking in such borderline cases needs to be confirmed. “At risk” candidates should therefore be reviewed by the most reliable markers.

All the discussions and decisions of the meeting should be recorded in a grade award report and used to help set assessments for the next session.

Feedback to schools

The focus of the assessment process is to provide valid grades to candidates, not to provide advice to schools on their teaching. Nevertheless, the IB is committed to supporting schools in improving standards. While it is not appropriate to spend resources providing individual feedback, subject reports should provide comprehensive information on the overall session performance. This has the double benefit of strengthening the support provided by summative assessment to classroom teaching and clarifying the workings of the assessment system for the schools, teachers and candidates who use it.

After each examination session, the examinations and their associated markschemes are made available for schools to purchase. In your principal examiner/chief examiner role, it is your responsibility to write the subject report using the information given to you by the examining team. The subject report covers all general aspects of candidate performance on each component, outlines where candidates performed well and where they seemed less capable, and makes recommendations for improving the preparation of candidates.

The Final Award Committee

The Final Award Committee (FAC) meets after all the grade award meetings have been held and just before the results are issued. This committee formally awards diplomas and certificates to those candidates who have met the requirements. The meeting is chaired by the Chair of the Examining Board and consists of a small number of other chief examiners and senior IB staff. An observer from a school is invited to attend.

The FAC has an academic honesty sub-committee that authorizes appropriate action on cases of alleged malpractice. The sub-committee comprises senior examiners, senior teachers from IB World Schools and IB staff. Each decision is made by a team comprising at least one representative from each group, a minimum of three people.

The committee also considers policy recommendations from IB staff on assessment arrangements for candidates with special educational needs. Cases of maladministration by schools that have abused deadlines and/or procedures are considered by the committee. For serious cases that involve a major threat to the security and integrity of the examinations, or for repeated maladministration, it is possible for a school’s authorization to be withdrawn.

Are you more generous when marking EURs?

The purposes of the enquiry upon results (EUR) service are to allow schools and candidates to highlight where they have concerns about the marking of assessments, and to offer transparency in how marks are awarded. It is not intended as a mechanism for candidates to receive a better outcome, and chief and principal examiners need to provide leadership to their examining teams to ensure that exactly the same standards are applied during EUR as during the initial marking period. All candidates must be treated fairly whether they have submitted an EUR request or not. A good question to ask is whether the work you are re-marking is of sufficient quality to deserve a higher grade.

Where problems with the marking are identified through the EUR process, that is, where marks are changed, in your role as chief/principal examiner you need to consider how these disagreements can be minimized in future sessions and how marking reliability can be increased to ensure correct outcomes for all candidates.

How do you make the assessments better for next time?

It is important that we learn from our experiences during the assessment cycle to improve both future examinations and the models we use to assess candidates. Chief and principal examiners need to provide clear guidance in identifying these improvements and support the IB in addressing them.

Assessment models are revised as part of the curriculum review process for each subject. Subjects are normally reviewed every seven years by a review group consisting of teachers, examiners, IB staff and external experts. The curriculum review process is consultative, with proposals being circulated to authorized Diploma Programme (DP) schools for comment as they are developed. Recommendations and proposals from the curriculum review groups are also submitted to the Diploma Review Committee (DRC) for consideration. This committee is responsible for the overall academic quality of the courses that make up the DP and approves proposed syllabuses and assessment models. The committee is particularly concerned with:

- the academic standard and comparability of different courses
- reducing overlap of subject content or objectives to a minimum and encouraging courses that complement each other instead
- monitoring the overall assessment burden on candidates, teachers and the IB, to ensure its manageability
- eliminating unnecessary duplication of assessment.

In the context of assessment, both the curriculum review groups and the DRC refer to the DP assessment policy, which defines the parameters under which assessment models are developed. Any issues with current assessments should be highlighted, and consideration should be given to improvements that can be made to current levels of validity, reliability and fairness.