International Baccalaureate®
Baccalauréat International
Bachillerato Internacional

Assessment principles and practices—Quality assessments in a digital age

International Baccalaureate®
Baccalauréat International
Bachillerato Internacional

Assessment principles and practices—Quality assessments in a digital age

Assessment principles and practices—Quality assessments in a digital age

Published July 2019

Published on behalf of the International Baccalaureate Organization, a not-for-profit educational foundation of 15 Route des Morillons, 1218 Le Grand-Saconnex, Geneva, Switzerland by the

International Baccalaureate Organization (UK) Ltd
Peterson House, Malthouse Avenue, Cardiff Gate
Cardiff, Wales CF23 8GL
United Kingdom
Website: ibo.org

© International Baccalaureate Organization 2019

The International Baccalaureate Organization (known as the IB) offers four high-quality and challenging educational programmes for a worldwide community of schools, aiming to create a better, more peaceful world. This publication is one of a range of materials produced to support these programmes.

The IB may use a variety of sources in its work and checks information to verify accuracy and authenticity, particularly when using community-based knowledge sources such as Wikipedia. The IB respects the principles of intellectual property and makes strenuous efforts to identify and obtain permission before publication from rights holders of all copyright material used. The IB is grateful for permissions received for material used in this publication and will be pleased to correct any errors or omissions at the earliest opportunity.

All rights reserved. No part of this publication may be reproduced, stored in a retrieval system, or transmitted, in any form or by any means, without the IB's prior written permission, or as expressly permitted by the Rules for use of IB Intellectual Property.

IB merchandise and publications can be purchased through the IB Store (email: sales@ibo.org). Any commercial use of IB publications (whether fee-covered or commercial) by third parties acting in the IB's ecosystem without a formal relationship with the IB (including but not limited to tutoring organizations, professional development providers, educational publishers and operators of curriculum mapping or teacher resource digital platforms etc) is prohibited and requires a subsequent written license from the IB. License requests should be sent to copyright@ibo.org. More information can be obtained on the IB public website.

International Baccalaureate, **Baccalauréat International**, **Bachillerato Internacional** and IB logos are registered trademarks of the International Baccalaureate Organization.

IB mission statement

The International Baccalaureate aims to develop inquiring, knowledgeable and caring young people who help to create a better and more peaceful world through intercultural understanding and respect.

To this end the organization works with schools, governments and international organizations to develop challenging programmes of international education and rigorous assessment.

These programmes encourage students across the world to become active, compassionate and lifelong learners who understand that other people, with their differences, can also be right.

IB learner profile

The aim of all IB programmes is to develop internationally minded people who, recognizing their common humanity and shared guardianship of the planet, help to create a better and more peaceful world.

As IB learners we strive to be:

INQUIRERS
We nurture our curiosity, developing skills for inquiry and research. We know how to learn independently and with others. We learn with enthusiasm and sustain our love of learning throughout life.

KNOWLEDGEABLE
We develop and use conceptual understanding, exploring knowledge across a range of disciplines. We engage with issues and ideas that have local and global significance.

THINKERS
We use critical and creative thinking skills to analyse and take responsible action on complex problems. We exercise initiative in making reasoned, ethical decisions.

COMMUNICATORS
We express ourselves confidently and creatively in more than one language and in many ways. We collaborate effectively, listening carefully to the perspectives of other individuals and groups.

PRINCIPLED
We act with integrity and honesty, with a strong sense of fairness and justice, and with respect for the dignity and rights of people everywhere. We take responsibility for our actions and their consequences.

OPEN-MINDED
We critically appreciate our own cultures and personal histories, as well as the values and traditions of others. We seek and evaluate a range of points of view, and we are willing to grow from the experience.

CARING
We show empathy, compassion and respect. We have a commitment to service, and we act to make a positive difference in the lives of others and in the world around us.

RISK-TAKERS
We approach uncertainty with forethought and determination; we work independently and cooperatively to explore new ideas and innovative strategies. We are resourceful and resilient in the face of challenges and change.

BALANCED
We understand the importance of balancing different aspects of our lives—intellectual, physical, and emotional—to achieve well-being for ourselves and others. We recognize our interdependence with other people and with the world in which we live.

REFLECTIVE
We thoughtfully consider the world and our own ideas and experience. We work to understand our strengths and weaknesses in order to support our learning and personal development.

The IB learner profile represents 10 attributes valued by IB World Schools. We believe these attributes, and others like them, can help individuals and groups become responsible members of local, national and global communities.

International Baccalaureate®
Baccalauréat International
Bachillerato Internacional

© International Baccalaureate Organization 2017
International Baccalaureate® | Baccalauréat International® | Bachillerato Internacional®

Contents

Introduction and overview 1

Introduction 1
Which IB programmes does it cover? 4
Why should you read *Assessment principles and practice*? 5
Who is this resource for? 6
Using this resource—different routes through the document 7
What does "quality assessment in a digital age" mean? 12
How does this resource relate to other IB resources? 13
Assessment using technology 15
What are assessment principles and practices? 19
Language of assessment 20
Emerging terms for eAssessment 22

Section A—Principles of assessment 24

What is assessment about? 24
Fit for purpose? Validity 29
Elements of the validity chain 35
Defining standards 51
Describing success—candidate achievement for summative assessment 56
Marking assessments 60
What is a good assessment? 65
IB's principles of assessment 77

Section B—IB assessment practices 78

What do we mean by a practice? 78
Reporting candidate achievement 79
Assessment process: Roles and responsibilities 88
Integrity of the assessment 92
Fairness for all—meeting candidates' needs 95
The assessment cycle 99
Examination paper preparation—development and quality 102
Examinations 107
Standard setting—Preparing examiners for marking 110
Marking 114
Moderation 126
Grade awarding (and aggregation) 133
"At risk" based quality checks 144
The final award committee 145

Preparation for release of results 147
Fairness for all—meeting candidates' needs 151
Enquiries upon results (EUR), appeals and general feedback 155
Setting next year's assessments 159
Feedback to schools 162

Section C—IB programme-specific processes 164

What are programme-specific processes? 164
Elements common to all programmes 166
Student competencies and the learner profile 168
IB Diploma Programme 174
IB Career-related Programme 180
IB Middle Years Programme 183
IB Primary Years Programme 187

Annexes 189

Annex 1: Moderation of internal assessment 189
Annex 2: Roadmap for creating a validity argument 195
Bibliography 197
Glossary 200

Introduction

This document is intended to explain the principles the International Baccalaureate (IB) has adopted to make sure that the assessment we undertake is meaningful, fair and in the best interest of the students involved.

It is written with teachers in mind, but should be accessible to students, examiners, stakeholders and other interested parties.

eAssessment is about being able to assess what is important, not just what is possible with paper and pen, and doing so in the medium that the current generation of students is most familiar with.

> Everything that can be counted does not necessarily count; everything that counts cannot necessarily be counted.
>
> (Albert Einstein/William Cameron 1963)

> Clearly, if the other criteria are less reliable than the examinations, greater reliance on them will lead to less reliable selection decisions.
>
> (Cresswell 1986: 37–54)

These two quotes indicate the scope of the challenge that we face with assessment. Many of the objectives for an International Baccalaureate (IB) education are not easy to assess, but without detailed assessments of our learners, important decisions that will affect their lives will be made on less fair and reliable grounds.

IB programmes are taught in over 140 countries by schools representing a wide variety of educational contexts and traditions. In some of these contexts, the philosophy and approaches adopted by the IB in assessing their students will seem familiar, while to others, the system might seem mysterious and obscure.

Such clarity is even more important during a period of change, and the impact of technology on education, including assessment, will continue to be felt over the next decade. We strongly believe that technology should support assessment and the move towards computerized on-screen examinations will not change our principles; but it may open up new possibilities in turning these principles into practices. More details on this can be found in the section on "Assessment using technology".

We believe that it is important that everyone in the IB community understands how our external assessment process works, what its strengths and limitations are, and the reasons why decisions are taken. Increased transparency can only lead to better understanding and ultimately a better education for our students. By using the opportunities offered by on-screen resources, we hope to provide teachers with a clear guide that is accessible but also contains the depth of information they need to understand IB assessment.

Figure 1

All assessments are a balance

All assessments are a balance between conflicting demands and many concerns about testing processes fail to take this into account.

An example might be the tension between reducing the assessments burden and the risk of candidates only having one opportunity to show what they can do.

For more information on this see the section on "Fit for purpose? Validity".

Need to balance between conflicting priorities

Another aspect of balance is the fact that the focus of the IB is to develop students through a holistic programme of study, and we must reflect this in our approach to assessment. This means we should make decisions about the impact on the overall programme, not a narrow focus on one subject, discipline or assessment.

In order to understand the nature of the IB assessment philosophy and operation, it is necessary to provide some background on the historical and theoretical development of assessment practice. Many significant issues are only briefly touched upon, but it is important to highlight them as they have a significant impact on current practice. For readers who wish to find out more, the academic papers quoted in the text will make a suitable starting point for further investigation.

We started by recognizing the difficult task the IB sets itself; to focus on what is important to assess and not what is easy. This is perhaps most eloquently expressed by Alec Peterson, the first Director General of the IB:

> What is needed is a process of assessment which is as valid as possible, in the sense that it really assesses the whole endowment and personality of the pupil in relation to the next stage of his life, but at the same time sufficiently reliable to assure pupils, parents, teachers, and receiving institutions that justice is being done. Yet such a process must not, by its backwash effect, distort good teaching, nor be too slow, nor absorb too much of our scarce educational resources.
>
> (Peterson 1971: 27–55)

We hope the rest of this resource explains how we believe we can deliver on this challenging objective and support the wider educational intentions of the IB in providing a world class experience for its students. If you have further queries which are beyond the scope of this resource please contact IB assessment staff, by emailing assessment@ibo.org.

Which IB programmes does it cover?

Previous versions of *Assessment principles and practices* have focused on the Diploma Programme (DP), but with the expansion of the full range of IB programmes, it is appropriate to explain in detail the IB philosophy on assessment that is applied across all areas: DP, Career-related Programme (CP), Middle Years Programme (MYP) and Primary Years Programme (PYP). While we will look in depth at the reasons behind the IB's assessment principles we will not provide much detail on their implementation in individual programmes. For this you should refer to the *From principles into practice* resource for the MYP, DP or CP.

This document focuses particularly on external IB summative assessment. Currently the DP, CP and MYP offer IB-run external summative assessment.

While assessment is an important element of the PYP, it is not appropriate for the programme to have any IB external summative assessment. For more insight into the specific philosophy and details of assessment in PYP refer to the assessment section of the PYP on the programme resource centre.

Our principles of assessment discuss a wide range of purposes for assessment, including formative assessment, which lead to our practices for the IB-assessed components. These principles will be of interest to all programmes.

Why should you read *Assessment principles and practice*?

Because assessment results have an impact on students' lives

The majority of the content of this document refers to the way in which the IB assesses candidates to award DP, CP and the optional MYP outcomes, which are then used by students to progress into further education or work.

Like everything else, assessment is only a tool which can lead to positive or negative outcomes for those being evaluated with them. As a teacher, parent or student, you are involved in assessments and should understand the strengths, weaknesses and decisions that those offering and using assessment need to make. Below are some of the common questions and comments we receive from teachers and how this document will help answer them.

"Why would the IB do that?"

This is the heartfelt question that we often hear from teachers and students when faced with external assessment. By reading this document you should understand the principles that drive our assessment practices.

"It's not fair, sir, I needed a higher grade!"

For students, a great deal will depend on their examination results, such as university entrance and future career, and they need to understand why decisions are made. Assessment is all about balancing conflicting and competing demands and this document will help you to explain to students the wider implications of bending the rules in their case, and maybe even convince them that their grade was "fair".

"I had not thought about it like that before."

Even for teachers who have worked in education for many years, external assessment can often seem a mysterious and opaque process. No part of an education, particularly an IB education, should be "done to" someone, and so we need to be able to explain to students how and why they receive the results they have.

"I only want to use assessments to inform better teaching."

An understanding of what makes good assessment, and what decisions need to be made, is important even if you are intending to use assessment for formative purposes. Poor quality assessments will lead to poor quality outcomes, and this is equally true if the outcome is to support learning or some other internal school purpose as if it is intended for selection or formal recognition. This document will help understand how the IB meets its objectives to provide high quality assessment to support its educational goals.

Who is this resource for?

We have assumed that the reader is a teacher who is familiar with the IB and its assessments but has not formally studied the theory of assessment. We believe that this means that the resource will be accessible to the widest possible audience.

It is likely to be of interest to students, parents, head teachers and IB coordinators and a glossary and links to other documents are provided in order to support a full understanding of all the processes involved.

In a similar fashion, the sections on the principles behind the IB's assessment practices will be of interest in understanding why we take the approach we do. The overview sections will be particularly helpful for examiners to place their role within the wider context of IB assessment.

Finally, university admissions officers and other stakeholders will be able to gain an understanding of what the IB expects to achieve from its assessments and what student outcomes represent.

Using this resource—different routes through the document

- The range of topics in *Assessment principles and practices* is so that it is helpful to provide a framework to guide readers to sections which are of particular interest to them.
- The **topic questions** below can act as a starting point for learning about those aspects that are most important to the reader.

This resource has been designed to have many routes through, depending on the interests and needs of the reader. The following table of contents is intended only as a reference list of all the sections rather than a suggested order in which they should be read.

What are these topic questions?

Assessment principles and practice is intended as a comprehensive overview of the way that the IB approaches assessment. Many teachers will not want this breadth and are looking for specific information about part of the process. To help do this we have prepared a number of separate lists of sections which concentrate on particular questions that teachers may have.

What difference will on-screen assessment make to IB assessment?

Intended for teachers who are familiar with IB assessment but are interested in how the IB will handle them, move to on-screen assessments.

Typical questions include:

- What is on-screen assessment?
- Will it be marked by computer?
- Why will it be better?
- Will it be the same standard?

A list of sections that are likely to be of particular interest to you can be found here.

Assessment using technology	What does on-screen assessment look like?	Risks with on-screen assessment to avoid
Language of assessment	What are the key terms to understand?	Difference between a candidate and student
Assessments, examinations, tests and components	Marking and grading	Difference between a question and an item
Paper authors and examiners	Emerging terms for eAssessment	On-screen assessment
Response file and candidate response	Familiarization tool	Unsure what a term means?
What is a good assessment?	What does good on-screen assessment look like?	The assessment cycle
Impact of eAssessment on the assessment cycle	Moving to an on-screen assessment	On-screen assessments

How are assessments marked and grades awarded?

An explanation to teachers on what goes on once they send the candidates' work to the IB for marking and how grades are produced.

Typical questions include:

- Why are there different grade boundaries this year?
- Why can't I have a different examiner?
- Who marks the scripts?
- How are the examiners checked?
- How can I appeal?
- I would have given the mark to the candidate—why is this fair?
- The markscheme indicates that the candidate's answer is worth a mark—how can this be fair?
- Isn't a "grade-free classroom" the best approach for students?

A list of sections that are likely to be of particular interest to you can be found here.

Language of assessment	What are the key terms to understand?	Difference between a candidate and student
Assessments, examinations, tests and components	Marking and grading	Difference between a question and an item
Reporting candidate achievement	What do IB grades mean?	What is the difference between marks and grades?
The tyranny of grades—lesser of two evils	What is a successful examination session?	Achievement is more than just grades
Assessment process: Roles and responsibilities	Principal Examiner and Chief Examiner	Other examiner roles
The responsibility of IB staff	Examiner hierarchy	Elements of the validity chain
Standard setting—Preparing examiners for marking	Standardization meeting	Practice scripts
Qualification scripts	Seed scripts	Tolerances
Successful standard setting	Marking	What is marking—consistent examiner judgement
"Definite marks"	Marks and grades are not the same thing	(Basic principles of) on-screen marking
Different types of markschemes	Analytic markschemes	Holistic criteria— markbands
Additional support for examiners	Question item groups (QIG)	Quality model
Practice scripts	Qualification scripts	Seed scripts
Tolerances	Challenging and unusual scripts	School connections
Aggregation	Grade awarding (and aggregation)	What is grade awarding?
Judgmental and interpolated grade boundaries	Impact of eAssessment on grade award	Evidence used in grade award
Considering this year's cohort	Feedback on the assessment	Reviewing script evidence
Reviewing statistics on outcomes	Balancing the evidence	Grade descriptors
Fixed grade boundaries	Aggregation	Quality checks on grade awards and distribution reports

Awarding a programme certificate	Teacher observers	Principles of grade award
"At risk" based quality checks		

What is the IB's approach to assessment?

For university admissions staff, teachers and stakeholders who are interested in the more theoretical underpinnings of the IB assessments; how we define good quality assessments, what directs our decision making in setting processes and how we balance the conflicting demands of assessment.

Typical questions include:

- What is IB assessment all about? What makes it special?

A list of sections that are likely to be of particular interest to you can be found here.

What is assessment about?	What is "assessment"?	Formative, summative and assessment as learning—why are we doing assessment?
Backwash effect and learning	Marks and grades are not the same thing	Fit for purpose? Validity
What does validity mean?	What is valid? Assessment or use	Creating a validity argument
Maintaining validity	Benefits of on-screen to validity? Benefits of eAssessment to validity?	Elements of the validity chain
Balancing aspects of validity	Reliability	Consistent outcomes are not the same as the right outcome
Construct relevance and authenticity	Manageability	Fairness and bias
Comparability	IB's approach to validity	Defining standards
Three meanings of standards	Norm-referencing and criterion-referencing of performance standards	What is norm-referencing?
What is criterion-referencing?	Which approach does the IB use?	Maintaining standards
Describing success—student achievement for summative assessment	The tyranny of grades—lesser of two evils	Importance of professional judgment
Marking assessments	What do we mean by "marking"?	Alternative forms of marking
Marking and formative assessment	What is a good assessment?	Good assessment supports curricular goals
What is good predictability?	Good assessment uses a range of assessment tasks	The role of classroom-based assessment and internal assessment
Collaborative working versus individual marks	Good assessment considers the wider student competencies and higher-order thinking skills	Higher-order cognitive skills
Student competencies and the learner profile	International-mindedness and intercultural understanding	What does good on-screen assessment look like?
IB's principles of assessment	Reporting candidate achievement	What do IB grades mean?

What is the difference between marks and grades?	The tyranny of grades—lesser of two evils	What is a successful examination session?
Achievement is more than just grades		

How does the IB design and write exam papers?

Intended for teachers who want to understand how exam papers are crafted and to provide reassurance for those who have concerns about the quality of the papers.

Typical questions include:

- Who writes exam papers?
- How are they checked?
- Are they the same when in different languages?
- Do they really test the right thing?

A list of sections that are likely to be of particular interest to you can be found here.

What is assessment about?	What is "assessment"?	Formative, summative and assessment as learning—why are we doing assessment?
Backwash effect and learning	Marks and grades are not the same thing	Fit for purpose? Validity
What does validity mean?	What is valid? Assessment or use	What is a good assessment?
Good assessment supports curricular goals	What is good predictability?	Good assessment uses a range of assessment tasks
The role of classroom-based assessment and internal assessment	Roles in authoring examination papers	Examination paper preparation—development and quality
Rules underpinning the writing of the examination papers	Question banks	Overview of examination paper preparation
Process up to resources sign off	Process to content sign off	Process to layout sign off
Process to usability sign off	Quality control	Translations
Managing requests for modified papers	Moving to an on-screen assessment	

How does the IB moderate teacher-marked assessment?

For teachers and stakeholders who want to understand why the IB includes teacher assessment in its processes but then does not always accept those teacher judgments.

Typical questions include:

- Why are my marks changed?
- Why does my moderation factor change between years?

A list of sections that are likely to be of particular interest to you can be found here.

What is assessment about?	What is "assessment"?	Formative, summative and assessment as learning—why are we doing assessment?

Backwash effect and learning	Marks and grades are not the same thing	Moderation
What is moderation?	Selection of candidate work	Unusual and atypical work
Failing to find a moderation factor	Dynamic sampling	Previous system—The moderation hierarchy
Internal assessment (IA) feedback	Standard setting—Preparing examiners for marking	Standardization meeting
Practice scripts	Qualification scripts	Seed scripts
Tolerances		

What does "quality assessment in a digital age" mean?

Technology is increasingly becoming part of all our lives. Digital communication forms such as messaging via social media are more common means of talking to people than writing letters, and in many jobs a computer is an essential tool for getting the work done. Computers and wider technology are also having a positive impact on teaching and learning, and the IB is committed to helping its schools make best use of them in the classroom.

For many students, writing essays and research is now done on a computer, and writing on paper for two or three hours has become an unfamiliar task—our exams should represent an opportunity for candidates to show what they understand, rather than being a unique experience which they need to master.

Including the line "quality assessment in a digital age" in the title of this resource recognizes that, if they are to remain relevant, our assessments need to be delivered in a way that students are familiar and comfortable with. More importantly, it recognizes that, when making the transition, we need to remain faithful to the underlying principles on which the IB is built. This means that the technology is driven by the assessment needs, and never the other way around.

Most significantly, talking about assessment in the digital age helps us to focus on making our assessments even better. It is not simply about preventing any negative impacts in moving exams and assessment done in the classroom (internal assessments) to an electronic format, but about using the technology to overcome the limitations the IB has been wrestling with for the past 50 years by relying on paper-based exams and coursework.

Using technology in assessment provides the opportunity to create more valid assessments and the tag line underlines our enthusiasm for this process, while reinforcing our commitment to the principles that underpin all IB assessment.

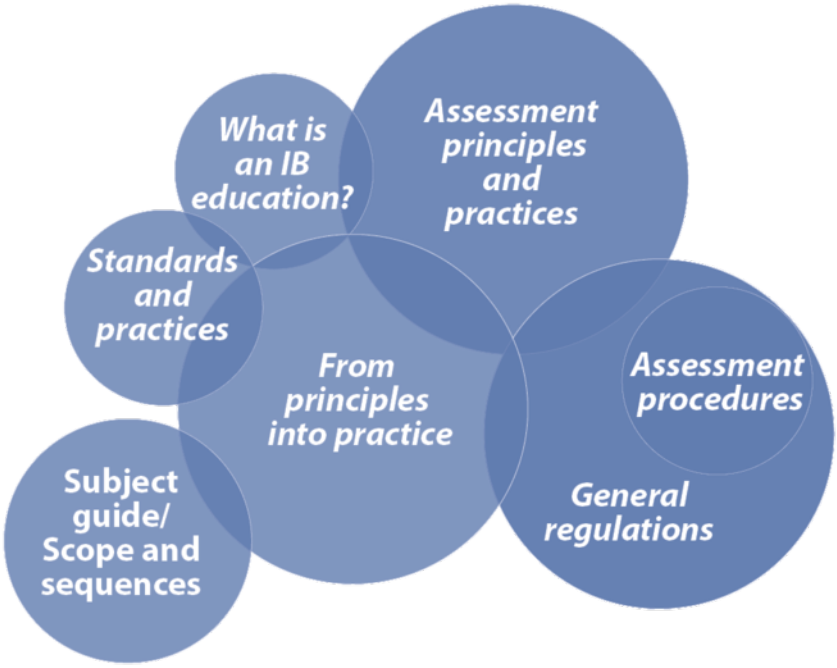
How does this resource relate to other IB resources?

This resource is one of several published by the IB to explain our approach to education. It can often be confusing to understand how these key resources relate to each other.

The diagram below illustrates this relationship while the table sets out the different purposes of several key IB resources.

Figure 2

Relationship between key IB publications



Key IB publications—covering all programmes	
Assessment principles and practices—Quality assessments in a digital age	Explains the overarching approach that the IB takes to assessment and how we intend to apply this in practice. It focuses on the summative assessment (formal testing) element of an IB education.
What is an IB education?	The aim of this document is to communicate clearly what lies at the heart of an IB education, explaining the ideals that underpin all IB programmes. By describing the IB's educational philosophy, the document also offers support for schools on their

Key IB publications—covering all programmes	
	journey from adopting the IB through many years of being an IB World School.
Programme standards and practices	Provides a set of criteria against which both schools and the IB can evaluate success in the implementation of the four programmes. It contains programme standards (general requisites for schools to implement any IB programme), practices (further definitions of the standards which are common to all programmes) and requirements (specific to an individual programme).

Key IB publications—programme-specific	
From principles into practice	Focuses on teaching and learning in the context of a particular IB programme. It also explains the requirements of the programme.
Subject guides (DP and MYP only)	Contain detailed information on one subject, course or discipline, for example, geography or visual arts. This includes setting out the aims, objectives, syllabus, and criteria for internal assessment in the particular subject. They usually also provide additional subject-specific guidance for teaching and learning.
Scope and sequence documents (PYP only)	The suite of PYP *Scope and sequence* documents offers examples of how to document curriculum expectations for each subject area. Their purpose is to: provide a tool to inform the whole school community about teaching and learning in each subject area; make transparent the essential elements of the PYP in the context of the subject areas; clarify the role of the subject areas in a transdisciplinary programme.
Assessment procedures	Sets out the rules, regulations and specific processes that must be followed in delivering the particular IB programme assessment.
General regulations	The regulations that underpin the rules we expect IB World Schools to follow.
Conduct of examinations (DP and MYP only—CP uses the DP version)	Informs coordinators and invigilators of the regulations concerning the administration and conduct of the programme examinations. A copy of this document must be available in every examination room.

In addition to these high level resources, the IB provides a range of other resources which provide focused guidance to teachers, coordinators and other stakeholders, which are based on the principles set out in these key publications. These can be found on the IB website and the programme resource centre.

Assessment using technology

- Students use technology widely, for example, in their social life, for their learning and to write their essays. Why do we still expect students to write their examinations by hand?
- The IB will use technology to deliver higher quality and more meaningful assessment. Technology will not drive the choice of approaches to assessment.
- It is important to separate the impact of technology to support expert examiners (e-marking) and using technology to create meaningful assessment for students (on-screen assessment (both for examinations and internal assessment) and ePortfolios).
- On-screen does not mean online: our assessments are designed to not require an internet connection in order to be taken. After they are taken the electronic files need to be sent to the IB using the internet.
- The technology that students use for assessment must be familiar to them from their classrooms. Generally, the use of technology in teaching pedagogy is better developed than in assessment.

What does on-screen assessment look like?

On-screen assessment is literally just doing an examination on a computer. It does not mean multiple choice or drag-and-drop (although we can include those if appropriate). The starting point has been to have our current style of examination papers on a computer. From there we will start to improve the assessments by including interactive questions that make the examination more authentic or more accessible for students. In the MYP, we have already started to explore the potential offered by on-screen assessments while in the DP we will start more slowly. The video below shows examples of the kind of questions on-screen assessments can offer.

What is eAssessment?

In practical terms, on-screen assessments mean each candidate needs their own computer in an examination. However, they do not need to be connected to the internet as the examinations are loaded on to the computers beforehand. The current model employed in the MYP is that the examination software "locks down" the computer while it is running and records any occasions where the software is interrupted for any reason (the invigilator can restart in cases of genuine disruption).

There are significant advantages to on-screen assessment for testing what we really want to test, rather than just what is possible with a paper examination. These are described in detail in the section on the benefits to validity but simple advantages include being able to offer video material, allowing candidates to rotate and move diagrams and to use common word processing tools when writing essays. The section on "Fairness for all" explains how on-screen assessment can also support accessibility, by giving the candidate control over colours and font size as well as tools like screen readers.

Finally, on-screen assessment will not remain static. As technology develops we will be able to develop our assessments to match. Currently, we are only considering students using keyboards and mice or track pads, but within a few years, touch screen technology may well become the norm and open up more possibilities. Looking much further ahead, what role could virtual reality play in certain assessments? It is possible to imagine foreign language assessments being more authentic in such an environment, and science fiction films provide an image for how technical tasks (science questions) could be managed through virtual reality laboratories.

Teaching using technology

Technology is currently changing the way students are taught, whether it is the use of interactive whiteboards in the classroom or massive open online courses (MOOCs) bringing teaching to a wider audience. Most students produce their essays on a computer rather than writing them by hand.

The move away from handwriting to using computers actually creates a worrying disconnect between assessment and teaching practices. For many students, the experience of spending two or three hours writing with a pen is almost unique to taking examinations and this significantly undermines the authenticity of the assessments.

The other side of this is that, with the introduction of on-screen assessments, we must make sure that candidates are familiar with the style of the assessments so that using the interface is not a barrier to them showing their understanding of the subject. While this can be achieved through familiarization tools and mock examinations, the ideal solution is for the students to be using the same tools during teaching.

What is e-marking?

E-marking is the term used for how we mark student work using computers to display the scripts and record the marks. In general, examiners are no longer sent the paper copies of examination papers or internal assessment work: this work is scanned into an electronic format and examiners can access them via the internet through specialist marking software.

Figure 3

IB examiners using the e-marking software to review candidate work during a grade award meeting

This has a number of major advantages for the speed and quality of marking.

- The IB always has the candidate work and if for any reason it needs to be looked at by a second examiner then it is instantly available rather than needing to be posted around the world.
- We can anonymize the work to reduce the chance of bias in examiner marking.
- We can implement a more rigorous quality control process as we can check marking standards during marking.
- Students from one school can be randomly shared between all examiners rather than all being marked by one examiner. This allows us to carry out additional quality checking processes.

E-marking has been used by the IB since 2010, and does not require students to submit digital work. Currently, most examination papers are handwritten and then scanned into a digital form.

Risks with on-screen assessment to avoid

In developing on-screen assessment, we are aware of a number of important challenges that must be overcome in order to ensure fairness to students and schools. The following list highlights only a few of the key risks we are managing. More detailed discussions will be included in individual sections of this document and in other IB publications dealing with eAssessment.

Burden on schools

We recognize that for students today the use of technology is an integrated part of their world and so it would be anachronistic to expect them to write on paper. However, we also recognize that schools need time to change their processes and that on-screen examinations require different arrangements from paper examinations. We will balance the introduction of on-screen assessments across the DP and CP to all schools to keep pace with technological advances and not place unreasonable expectations on schools to adapt, whatever part of the world they are in.

Risk of technology failure

There is no acceptable level of technology failure for students taking examinations. Every student must have a smooth and uninterrupted experience. Our commitment to delivering this is built on two principles. Firstly, that the assessment can be taken on an isolated machine without it needing to be connected to the internet. The second is that we have a comprehensive testing/compliance process which is available to the school before the day of examinations. Our experience is that most issues are as a result of school infrastructure or setting, and these can be identified and resolved well in advance of examinations.

Security issues

The assessments need to be at least as secure as paper examinations. Our assessments are designed to "lock down" all other functionality of the computer while they are being taken, and are encrypted and password protected.

While this remains an area we will continue to strengthen as more sophisticated approaches are developed by industry, we also are aware that current paper-based processes are subject to their own security risks.

Technology for the sake of technology

The IB is committed to using technology to improve the quality of what can be assessed, not using it for the sake of having it. We will continue to articulate how we have added value to the validity of our assessment, and our paper development process will reflect this to prevent using technology for the sake of it.

Bias against certain groups of students

We recognize that certain students are going to perform better as a result of being able to use computers to answer their assessments. However, we also recognize that paper examinations are not free of bias: students who find it difficult to write with a pen for a long time, or who are slow at writing, are already disadvantaged.

As a principle, in neither case (paper or on-screen) are we intending to measure the candidates' skills in using a pen or a computer and we will design assessment tasks that do not give advantage to those who can type or write quickly.

The IB is paying close attention to the published research on such "device effects" to ensure we can meet this principle.

Changing standards—up or down

IB grades have a meaning, and we will protect this meaning through the expert judgment of our senior examiners supported by outcomes data. As an example, it may be easier to create a better structured essay

when you can cut and paste paragraphs as you go along. We will make sure that this is taken into account when setting grade boundaries, so a candidate equally comfortable with paper and on-screen will obtain the same outcome—one which reflects their understanding and analytic ability in the subject.

As a result of being able to test new traits through on-screen assessment, and traits that we have always valued but not been able previously to assess, we may need to shift what we expect students to demonstrate in the assessments (see the section on "Defining standards"), but we will only do so in a way that upholds the long-standing goals of the IB and reflects the qualities described in the current grade descriptors.

Barriers to schools offering IB programmes

Linking to the point relating to the burden on schools, we will work with the IB community to have timescales to allow all those schools who want to take advantage of the opportunities offered by technology for better quality assessment to be able to do so. The IB believes that the potential of technology should be utilized intelligently to improve the quality of teaching, learning—and therefore assessment.

What are assessment principles and practices?

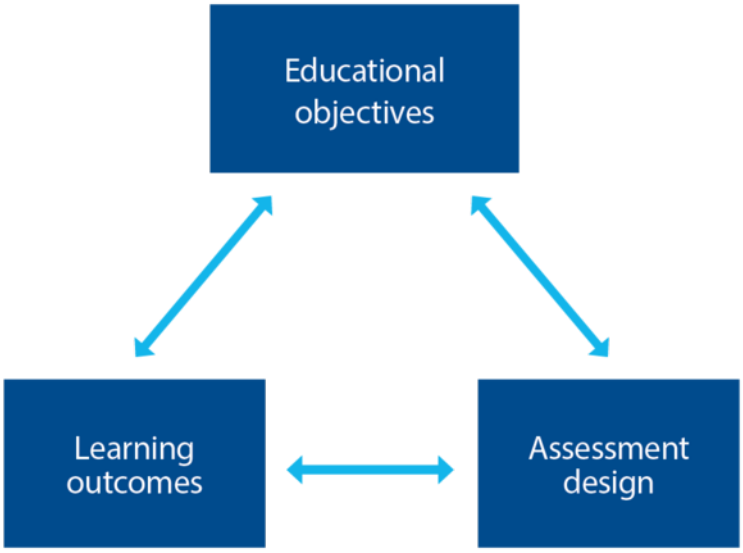
Assessment principles are what we think are important in creating, delivering, marking and grading qualifications and assessments. They come from what we think is important about an IB education and the most important principle is that assessments should support education, not distort it.

Assessment practices are the way in which we deliver our principles in a meaningful and practical way. They take into account the conflicting demands and practical limitations of working in the world while maintaining the IB philosophy of being principled.

The assessment practices are still described at a high level and should be implemented appropriately within the context of individual subjects, disciplines and programmes. They explain broadly what needs to be done, but without detail on the day-to-day process and the differences in implementation in individual subjects.

Figure 4

Relationship between objectives, outcome and design



Language of assessment

- Like most areas, assessment has its own language and set of terms that it is helpful to be aware of when talking about assessment.
- The introduction of on-screen assessment will also lead to a change in some of the language used, for example, if "paper" is still meaningful.

What are the key terms to understand?

Like many areas of life, including teaching, assessment has its own language and collection of terminology. We appreciate that this can make it difficult to engage in discussion about assessment and the IB is committed to the use of plain English and minimizing the use of jargon.

Despite this commitment, there are some words and concepts it would be useful to understand before reading this guide.

Difference between a candidate and a student

A candidate is someone who is taking our assessment. A student is someone who is studying our courses. Students become candidates when they take the assessment.

This difference is generally only important when dealing with data where numbers or outcomes may be different between the two groups.

Assessments, examinations, tests and components

Assessments are any of the tasks completed by candidates to demonstrate their ability. Examinations are one form of assessment, involving candidates completing IB-set questions under tightly controlled conditions.

In this document, we will use test in a generic sense to mean examination, although it is sometimes given a specific meaning in academic papers on assessment.

Sometimes the overall assessment of a candidate will be broken down into several separate pieces taken at different times. These are called components of the assessment or just components. Examples of components might be oral examinations, (examination) paper 1, (examination) paper 2, or the internal assessment.

Marking and grading

Marks are given to reflect how much of a question the candidate has answered correctly and the allocation of marks is different for each question and examination. A grade describes how good the candidate's performance is, and should mean the same for every examination, year and subject. This is a really important distinction.

More details can be found in "What is the difference between marks and grades?"

Difference between a question and an item

A question is the general term for a discrete task, which may have several parts to lead the candidate through the problem.

An item, or question item, is each individual answer a candidate gives, so question 1 part c (1c) or question 7 part e section iii, 7 part e iii would be items.

Paper authors and examiners

A paper author is the person who creates the questions and the associated markscheme that will be used for the assessments. Paper authors are nearly always senior examiners, but it is a job requiring a different set of skills and not necessarily linked with the examiner role.

An examiner is the person who marks the candidate's work. There is a hierarchy in examiner teams, with the two highest roles in the hierarchy being a Principal Examiner (PE) for each component and Chief Examiner (CE) for a subject.

Senior examiners, led by their PE and CE, are responsible for recommending where the grade boundaries are set. This boundary setting should be done after all the marking has been completed.

For more information on the responsibilities of the different roles see the section on "Assessment process: Roles and responsibilities".

Emerging terms for eAssessment

The move to eAssessment means that the IB and wider assessment community need to start developing new words to describe what we are doing. While sometimes we can follow the trend to add an "e" in front of a familiar word, for example, ePapers for the electronic version of an examination paper, there are some other terms that are very different from their paper-based equivalent.

Of the list below, only on-screen assessment is widely used in this document, but you may encounter the other terms if you read more widely about our eAssessment.

On-screen assessment

On-screen assessment is any assessment which takes place with a computer. It covers the idea of computer-based examinations, but can also mean classroom-based assessment work if it is done primarily in a computer environment.

The key difference between an on-screen assessment and just using a computer for class work is that, in an on-screen assessment, you will be using carefully controlled software which gathers the assessment evidence. For example, candidates may be allowed to use the internet to investigate a topic; but in the on-screen assessment it is likely that it will either guide the search and/or record where students have looked so that their approach to searching the web can be evaluated.

The choice of the term on-screen has been chosen very carefully. It is not online, and in particular connection to the internet is not required for candidates while taking our on-screen examinations.

Response file and candidate response

In the paper-based world, we talk about candidate scripts, but when we start to use the full range of opportunities offered on eAssessment, this candidate work could also be multimedia. The process of producing the work is part of what is being assessed. The outcome of an eAssessment is therefore known as a candidate response.

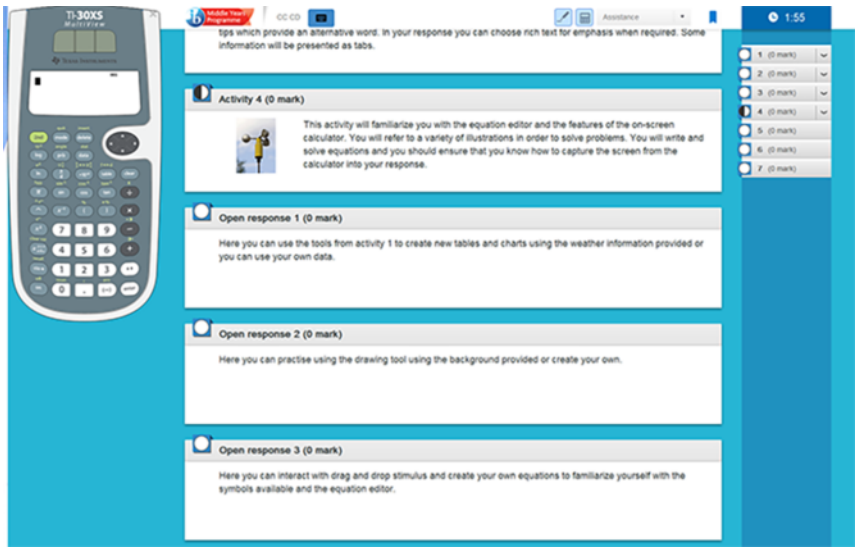
The response file is simply the computer file containing the candidate's response. These files are what need to be sent to the IB.

Familiarization tool

It is absolutely critical that candidates can use the computer effectively so they are not disadvantaged in their examinations. The familiarization tool is a content-free example of an on-screen assessment where candidates can practise using all the different features of the assessment.

Figure 5

Screenshot from the MYP familiarization tool



We expect candidates will also sit mock examinations using on-screen assessment, but they will be doing so in timed conditions and worry about their answers to the questions. In contrast, they can have unrestricted access to the familiarization tool to familiarize themselves with how the software works.

Unsure what a term means?

We recognize that IB assessments use a large number of terms which can be unfamiliar to teachers and students. Throughout all our documents, we will try to avoid confusion while remaining precise in what we are saying.

As part of this commitment, we will be careful in our use of abbreviations, making sure they are spelled out in full the first time they are used in each section. Similarly, we will make sure that important terms are explained the first time they are used, or that links are given to where they are explained.

If you come across any words you are not familiar with, this document has a full list of assessment terms in the glossary.

What is assessment about?

- Assessment can mean many different things.
- In education, assessment has often been divided into formative (assessment for learning) and summative (assessment of learning). Today, there is a strong movement for "assessment as learning".
- Assessment must be designed carefully to meet the purposes its results are used for. An excellent formative test may be very poor for measuring summatively.
- Assessment can influence teaching practices and must be designed so any "backwash effect" is positive.

What is "assessment"?

"Assessment" can mean any of the different ways in which student achievement can be gathered and evaluated. Common types of assessments include tests, examinations, extended practical work, projects, portfolios and oral work. Sometimes, assessments are carried out over a prolonged period, and at other times they take place over a few hours. Assessments will sometimes be judged by the student's teacher, while other times they are evaluated by an external examiner.

You will notice that we have used the terms evaluated and judged rather than talking about marking or grading. This is because there is an important distinction between these two concepts which is explored later.

Figure 6

Examples of different types of assessments

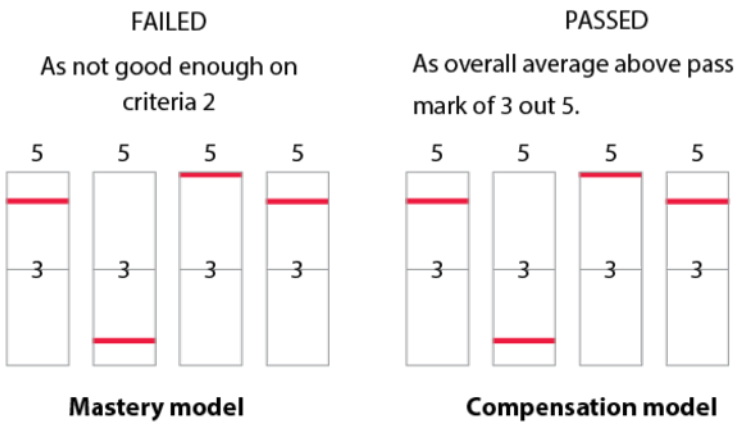
Assessment tasks that test how good (competent) a candidate is can use either a **compensation** model or a **mastery** model. The compensation model is used in most external exams, and it allows an excellent performance in one area to mitigate for a poor performance in another. As an example, imagine an exam

consisting of three questions marked out of 10. The pass mark of 15 could be achieved by a candidate scoring of 5, 5 and 5 (achieving around half of each question) or by a candidate scoring 10, 3 and 2. In the second case a perfect score in the first question compensates for poor scores in the other questions.

In contrast, a mastery model of assessment requires a minimum attainment (mastery) in each part of an assessment. In the example above, if a 5 was required in each question, the first candidate would pass, while the second candidate would fail, as would a third candidate who achieved 10, 10 and 4, a total of 24 marks overall but not the required 5 in question 3.

Figure 7

Comparison of mastery and compensation models



Mastery model assessments are often used in more vocational (workplace) settings where it is not appropriate to be really good at one element but poor at another. For example, when making a dress no degree of design expertise can compensate for not having basic sewing skills.

The IB employs a range of assessment tools, including examination papers intended to be taken at the end of the programme, and a variety of other assessment tasks (essays, research essays, written assignments, oral interviews, scientific and mathematical investigations, fieldwork projects and artistic performances) spread over different subjects and completed by candidates at various times under various conditions during their course.

Formative, summative and assessment as learning —why are we doing assessment?

Assessment can be used for a variety of different purposes. The intended purpose for a given assessment will have a major impact on how it is designed. Traditionally, there have been two broad reasons for doing assessment: formative and summative.

For formative assessment, the aim is to provide detailed feedback to teachers and their students on the nature of students' strengths and weaknesses, and to help develop their capabilities. Types of assessment such as direct interaction, for example a discussion, between teacher and student are particularly helpful here.

Vygotsky (1962) describes the teacher as being seen as a supporter rather than a director of learning and so should make use of assessment tasks and instruments that help the student work in what he refers to as the "zone of proximal development". This is the range of achievement between what the student can do on their own, and what the student can do with the support of the teacher.

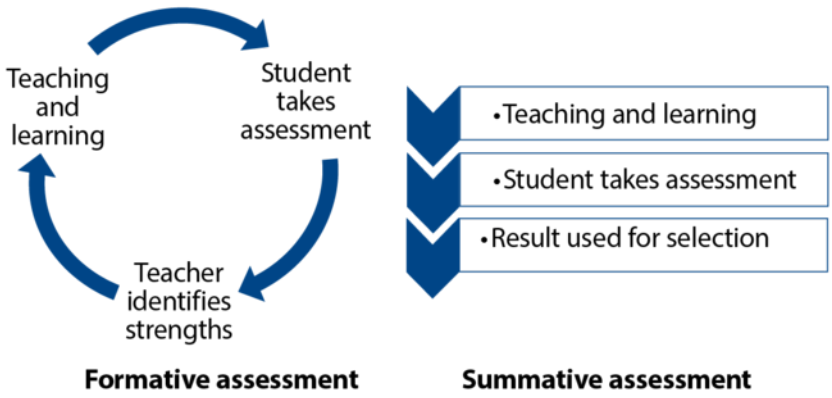
This concept of the notion of "scaffolding" was formed by Wood et al (1976), where the teacher provides the scaffold for the construction of learning but only the student can do the constructing. The intention of

the teacher must be to set formative assessments that are at just the right level of challenge for the student, and to keep adjusting that level as the student progresses.

In contrast, summative assessment focuses on measuring **what** the candidate can do, typically to demonstrate the completion of a training programme and/or readiness to progress to the next stage of education. While formative assessment is interested in **why** a student does something, summative assessment wants to know whether they did the **correct** thing. While this may seem less useful than the why question, consider the different purpose of summative assessment, which is to make a judgment about the candidate, not to inform future teaching. For those not convinced on the need for summative assessment, we suggest reading the section on "Describing success".

Figure 8

Two possible differences in how formative and summative assessments are used



In formative assessment, it is more important to identify correctly the knowledge, skills and understanding that students have not yet developed, rather than to measure accurately the level of each student's achievement. The balance between these concepts is called validity and is discussed in more detail in later sections. This balance between the student's attainment and the quality of feedback is reversed in summative assessment, where the outcomes of the assessment will be used to make decisions about the student, often around competitive selection for employment or educational opportunities, but also to support further teaching.

It is worth being aware that any analysis of different national assessment systems will quickly reveal a wide variety of assessment techniques and approaches. All of these systems have their strengths and weaknesses in relation to technical, resource and time considerations, and in their impact on the country's education system. Even if it were possible, in a given context, to start completely afresh in devising an assessment system, there is no universal best technical practice that could be adopted. Instead, the choices made in devising assessment systems inevitably reflect the values and priorities of the broader social context in which they are made. For more research in this area, see Cresswell (1996) and Broadfoot (1996).

It is also important to recognize that summative assessment is increasingly being used as a measure of the quality of teaching which adds a further dimension to why assessment is undertaken, for the benefit of the education system rather than the student.

Backwash effect and learning

> Nobody grew taller by being measured.
>
> (Meighan 2004)

> What is needed is a process of assessment which is as valid as possible. Yet such a process must not, by its backwash effect, distort good teaching, nor be too slow, nor absorb too much of our scarce educational resources.
>
> (Peterson 1971)

In this quote, Alec Peterson, the founder of the IB, recognizes the risk that the way in which assessment is carried out can influence the approach taken to learning. Surgenor (2010) neatly sums up why this can have both negative and positive implications.

In his 1971 paper, Snyder proposed that students create their own understanding of the curriculum based on implicit and explicit messages about what counts in assessment which is described as the "hidden curriculum". This can lead to understanding how to pass a subject but not understanding the subject itself. This is neatly summed up by the quote below.

> When I retook the exam I just concentrated on passing the exam. I got 96% and the guy couldn't understand why I failed the first time. I told him this time I just concentrated on passing the exam rather than understanding the subject. I still don't understand the subject so it defeated the objective in a way.
>
> (Gibbs 1992: 101)

This concept of backwash is also summed up in the often quoted adage "if it is not tested it will not be taught", highlighting that the assessment and teaching cannot be considered as independent of each other.

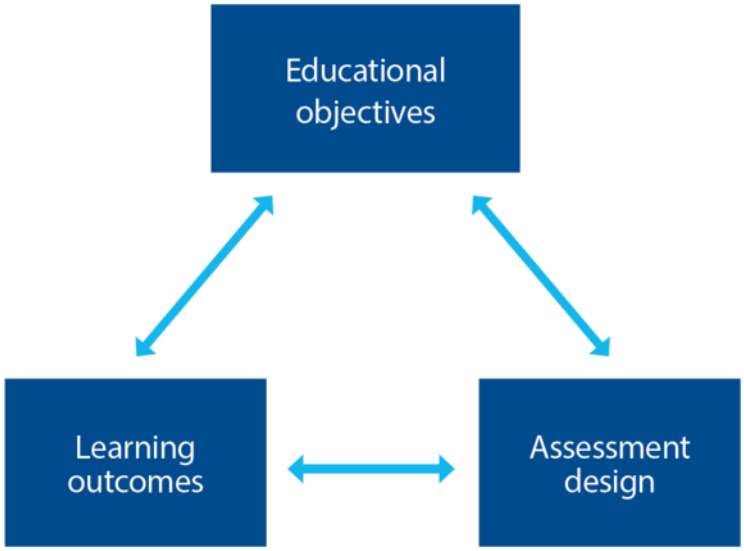
There are various ways in which an assessment can fail to support the educational objectives of which it is trying to get candidates to provide evidence.

The most likely is that the learning outcomes on which the assessment is based are not a good reflection of the original purpose of education. A vocational example would be a course to train elite athletes whose learning outcomes were around understanding the rules of athletics. Another common problem is that while an assessment covers every intended outcome, most of the testing focuses on tasks that are easy to test but not very important to the overall educational goals.

This interdependency of assessment and educational purpose can be expressed in the diagram below (from Furst's paradigm, 1958, in Frith and Macintosh, 1984). A disconnect between any two of the three elements will almost certainly lead to poor quality assessment.

Figure 9

Relationship between objectives, outcome and design



As set out in *What is an IB education?* we follow a constructivist approach to learning which offers an active role for the student and also recognizes the importance of context to effective learning (Murphy, 1999). If assessment is to support effective teaching and learning, then it must be designed around this constructivist learning theory. For more research relating to this concept see Black (1999); Shepard (1992); Wood (1998); and Lambert and Lines (2000).

Formative assessment has the most direct link to the way students learn, and is sometimes called assessment *for* learning while summative assessment is sometimes referred to as assessment *of* learning. This underestimates the major impact of summative assessment on what is actually learned in the classroom. All assessment should support appropriate learning. Summative assessment is not just an activity conducted after learning has taken place, but should be designed to have an integrated role in teaching and learning.

Marks and grades are not the same thing

An important aspect of carrying out, and using, summative assessments of candidates is to understand the difference between marking their work and grading their work.

- In marking, a candidate is given credit for the work they have produced against a markscheme or similar framework. This is an indication of the degree of the assessment task they got right. The mark itself has no other meaning.
- In deciding a grade, the examiner is making a judgment on the quality of the candidate's work against a defined standard which will take into account the difficulty of the task as well as the proportion of the task that was completed. The grade therefore has some meaning or relevance and is usually intended to be comparable with performances on other assessments.

Figure 10

Marking can be thought of as the quality (tastiness) of the food cooked, but the grade reflects the complexity of what they were trying to cook

It might be possible for a candidate to demonstrate a high grade from getting only a small proportion of a very difficult question correct, and be impossible to demonstrate the same grade by correctly answering many trivial questions.

As discussed in later sections, it is not necessary for the standard described by the grade to be explained by reference to what the candidate has attained, although this is the approach taken by the IB. There are other perfectly consistent and well-respected systems where the standard is based on how the candidate performs relative to peers.

In our assessments, the IB generally uses marks as an indication of overall performance (compensation model) and then looks at how candidates with this number of marks performed to determine a boundary point (grade boundary) where students with more than that number of marks are awarded a particular grade. This process is explained in more detail in the "IB assessment practices" section.

Fit for purpose? Validity

- Validity means asking if an assessment is fit for purpose. It is a complex concept with many aspects.
- Assessments are used for many purposes.
- An assessment may be valid for one purpose, but not another. For example, a test of spelling will not also measure fluency in a language.
- It can be argued that it is the purpose assessment outcomes are put to that are either valid or not, rather than the assessments themselves.
- In the IB, our first concern is whether the programme is valid, then whether elements of the programme (such as individual courses) are valid and finally whether an individual assessment is valid.

What does validity mean?

> Everybody is a genius. But if you judge a fish by its ability to climb a tree, it will spend its whole life believing that it is stupid.
>
> (Anonymous, although often credited to Albert Einstein)

What is the purpose of taking exams? What are exams for? These deceptively simple questions lie at the heart of what it means for an assessment to be "fit for purpose" or "measuring the right thing"; and often the various different answers to them are in conflict with each other. It is also the case that not all purposes of an education might be tested by any particular examination, or indeed some purposes might not be possible to test.

As an example, suppose that the following four possibilities were given as the purpose of an assessment in mathematics:

- to recognize what the student understands after their course of study
- a means of selection for further study or work
- an indication of future success
- to reinforce the teaching of the curricular goals of the programme.

Even these simple goals are difficult to reconcile with one another: if future success in mathematics depends more on calculus than on geometry, should we focus on calculus? How does this relate to the first objective of recognizing what the student knows (in geometry)?

These four possibilities are in no way definitive. Newton (2007) sets out a number of different examples of possible uses for assessment results.

• Social evaluation uses	• Licensing uses
• Formative uses	• School choice uses
• Student monitoring uses	• Institution monitoring uses
• Transfer uses	• Resource allocation uses
• Placement uses	• Organizational intervention uses
• Diagnosis uses	• Programme evaluation uses
• Guidance uses	• System monitoring uses
• Qualification uses	• Comparability uses
• Selection uses	

Faced with this daunting array of conflicting uses, it might be tempting to suggest that the assessment and exam grades cannot be fit for purpose and to avoid them. The problem with this approach is that there is still a need for people to make decisions and if assessment results are not available then they are likely to use other, less well-designed and understood ways of making these decisions. This is the point made in Mike Cresswell's quote.

> Clearly if the other criteria are less reliable than the examinations, greater reliance on them will lead to less reliable selection decisions.
>
> (Cresswell 1986: 37–54)

This concept of a qualification being "fit for purpose" is broadly what is meant by "validity".

The term validity is often used, but there is a long academic history of trying to define its meaning accurately (Newton 2012). As an added complication, there is a narrower concept of validity, often used together with "reliability", to mean an assessment is the extent to which it actually measures what it is stated to measure. An example would be not using a written exam to show the candidate can bake a cake. To avoid confusion in this resource we will refer to this second concept as construct relevance. See the "Construct relevance and authenticity" section for details.

Validity is ultimately a matter of judgment. The conflicting demands of an assessment alluded to above will be dealt with in more detail later, but there will always be a compromise between these demands. The decision that then needs to be made is whether the available evidence suggests that the assessment is sufficiently "fit for purpose" to be useful. Therefore, when talking about validity it is good practice to talk about the strength of this evidence or the "validity argument" being made rather than a simple yes or no. This idea is dealt with in more detail in the following quote.

> If the declaration of validity acts as a promise or guarantee, which provides users with a green light to interpret and use assessment outcomes as specified, this raises a final fundamental point which does not always receive the attention it deserves: without an explicit formulation of the validity claim users will not be able to interpret and use assessment outcomes appropriately.
>
> (Newton 2007: 149–170)

Finally, validity is not something that is achieved or not during the design of an assessment, but is continually developing during its life cycle. Similarly, the validity argument is not made at the start of the cycle, but is continually added to and refined during the process.

What is valid? Assessment or use

Consider the following situation.

A candidate has studied physics for two years and has achieved a grade 6. They took their exam in Spanish.

Does this grade mean that they are fluent in Spanish? Stretching the example to the point of absurdity, does it mean they are a good cook?

In this example, the examination that the candidate took may have been a good test of their understanding of physics, but it is being used as an indication of their understanding of Spanish. Is the assessment then valid?

A more subtle example could be whether a particular grade in mathematics indicates that the candidate is competent at arithmetic (which is only one aspect of the curriculum). This introduces the idea that it is not the assessment (or result) that is valid, but the purpose to which it is put. Newton (2012) expresses this as "the use of a given assessment procedure for a specific purpose (that is, to make a specific decision) is valid if its interpretive argument is sufficiently strong".

If it is the purpose the assessment is used for, and not the assessment itself which is valid, then what responsibility lies with the IB to show its exams are valid? Is it necessary to show validity for every possible use? Is it appropriate to define a limited number of uses for which the IB claims validity? Should we be explicit in stating the uses for which we do not claim validity?

We need a practical solution to this problem. Assessments are designed with a specific number of purposes for which the IB will make a "validity argument" to support. We will also consider the degree to which it meets other common uses which lie within the philosophy of the IB. If we decide that the arguments for

these two sets of purposes are sufficiently strong, then we will talk informally about the assessment being valid.

Creating a validity argument

Validity is not a simple objective concept but is a balance between competing issues. This means we cannot "prove" validity but only put together a compelling argument as to why the choices made have led to a meaningful course and assessment.

For similar reasons, the evidence for a validity argument should arise naturally out of the process of developing, monitoring and delivering a programme, and should reflect the discussions and decision that were reached. An example in assessment would be that the judgments behind what questions to ask in a particular paper will demonstrate that the curriculum has been covered appropriately.

The essential element of the validity argument is to have an appropriate structure that covers all the aspects we believe to be important. Within the IB this is represented by a series of questions for which evidence is gathered. These are listed in "Annex 2: Roadmap for creating a validity argument".

Maintaining validity

Validity is not something which is decided or "proved" when an assessment model is designed, but it develops with the assessment during its life cycle, and indeed beyond for as long as decisions are made based on its outcome.

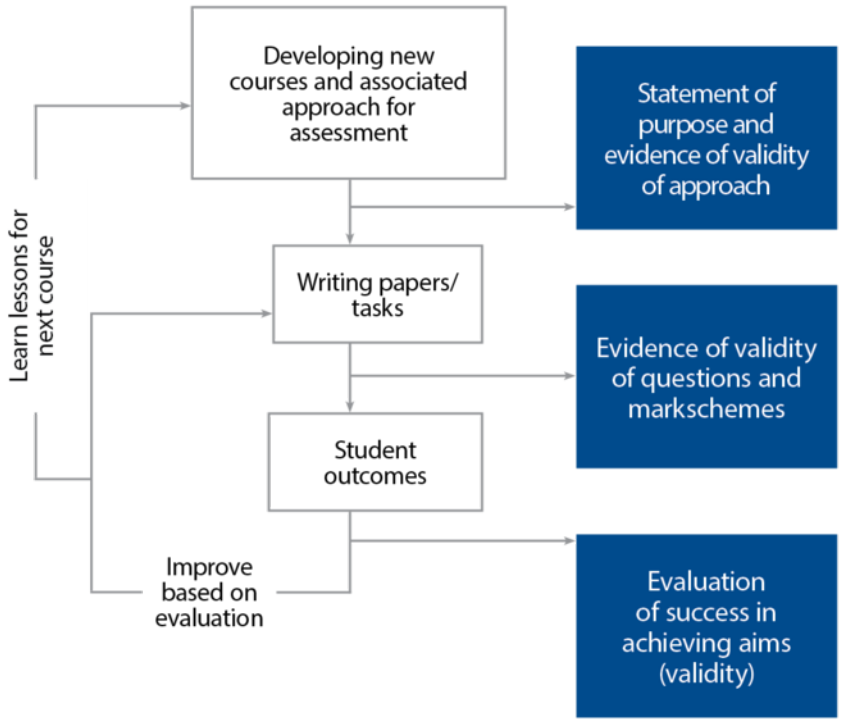
For the IB, different evidence to demonstrate the assessment is valid (that is, the validity argument) will be gathered many times throughout the life of a particular course.

When the course is developed or reviewed, then the discussions about the purpose of the course and how it should be assessed to meet this purpose will form the core of the validity argument, particularly the balance between construct relevance and the other aspects of validity. For every session, the writing of a new paper or assessment task will generate more evidence, especially fairness, comparability and construct relevance. Teacher feedback and student outcomes will provide an evaluation of how successful the assessment was in achieving its objective as well as providing evidence on the reliability of marking and comparability of grades. This information will then feed back into the development of the next set of assessments.

Finally, the evidence gathered from all the assessments for a particular course, and all the courses of study within a particular programme, provide the basis for decisions for the next review of the course/programme. The diagram below outlines this process.

Figure 11

Gathering evidence of validity throughout the life of an assessment



While the complete set of evidence for the validity argument can only exist once the assessment has been taken by the candidates, each stage provides the foundations for the next step in the process.

> Evaluating test validity is not a static, one-time event: it is a continuous process.
>
> (Sireci 2007: 477)

Benefits of on-screen to validity? Benefits of eAssessment to validity?

- Today's world is not paper-based: computers are part of every aspect of life.
- Paper examinations cannot provide moving images or allow meaningful interactions with the candidates.
- The versatility of computers allows candidates to make the visual and audio modifications they need to access the questions rather than needing to request special papers months in advance.
- Despite concerns, eAssessment offers more protection against academic misconduct and maladministration.

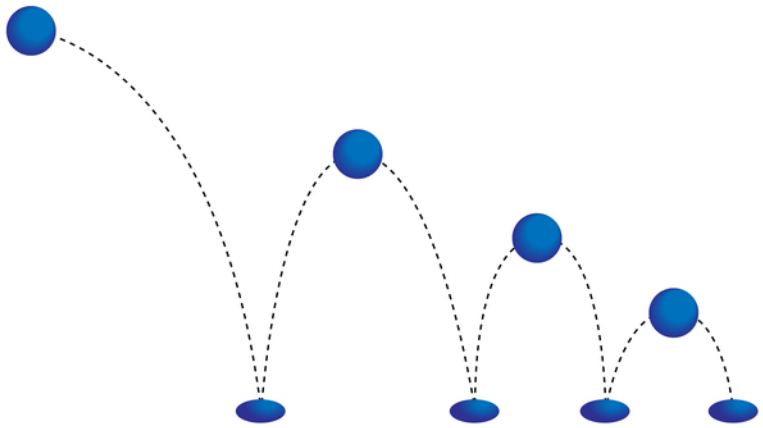
You are almost certainly reading this document on a computer as it is only available electronically. Email, texts and social media have become far more common than letters as a means of communication, and using a computer is a routine part of many jobs rather than something exceptional. If we are aiming for authentic assessments, we need to include the use of computers in our assessments.

Historically, assessment has been very limited in what it has been able to assess. Generally, within written examinations candidates have been limited to responding to a simple stimulus or question without any opportunity to manipulate or interact with the assessment. At the most basic level, on-screen assessments offer us the chance to include film and audio as stimuli, both of which can be played and paused by the individual candidate rather than relying on an invigilator playing material for the whole class. It also allows

for animated diagrams to be included where appropriate rather than a picture and a long description of how an object is moving.

Figure 12

The clarity of an eAssessment question about the trajectory of a bouncing ball could be improved by the use of a moving image rather than a static image, as shown here



As eAssessments become more sophisticated, it becomes possible to assess how the candidate interacts with a problem, responds to new information, or develops a simulation. Ultimately, it may be possible for an on-screen assessment to respond to prompts from the candidate, recreating the opportunities offered in teacher-run assessments, such as an oral examination, but without the problems of different candidates having different teachers.

The key benefit of eAssessment is that it allows us to assess what we actually want to test, rather than being limited to what can be assessed in a paper examination.

Another aspect of validity is minimizing bias, particularly for those with assessment access requirements. The IB regularly processes requests to produce examination papers with different fonts or different coloured paper. There are also requests for examination papers produced in braille.

While on-screen assessments cannot remove these barriers, it does allow us to empower the candidate to select the font size and colour which best suits their need. There are also major accessibility benefits for candidates with the use of screen readers. While some candidates may not be able to use the computer to address their access requirements, on-screen assessment will allow access to a wider range of candidates without the need for additional support, which can still be used where appropriate.

Figure 13

Where might eAssessments be vulnerable?

One of the common concerns with eAssessment is that it is less secure than paper examinations. In reality, on-screen assessments are protected from some of the risks paper examinations face, but are susceptible to others.

One of the biggest benefits is that the eAssessment can be sent electronically around the world, and remain "locked" until the passcode is given to candidates on the day of the examination. This compares favourably with the risks of sending papers, which can be read by anyone, and then requiring schools to keep them secure for several days before the examination. Even if security at the school is breached, the culprits would still need to break into the eAssessments themselves which have been designed to prevent such hacking. If this succeeds the culprits are still only at the same point they would have been if they had stolen paper examinations, and probably with less time as eAssessments can be delivered to the schools closer to the exam date.

We recognize that there is a challenge during the examination itself as the candidates have access to a computer which, in theory, could connect more easily to unauthorized notes and so on than the traditional methods of smuggling notes into the examination hall. IB eAssessments are designed to operate in "kiosk" mode which prevents access to any other program while they are running. Further, the ability to record how the candidates answer questions, as well as their final answers, provides opportunities to identify behaviour associated with cheating.

In summary, eAssessment does create new challenges around examination security, but it also removes some of the old opportunities for cheating.

Elements of the validity chain

- There are many different aspects of validity. If you think of them as a chain, then if one link is broken so is the whole assessment.
- The IB generally focuses on five elements in validity: reliability, construct relevance and authenticity, manageability, fairness and bias, and comparability.
- These five elements are often in tension with each other. In considering whether an assessment is valid you need to consider the main purpose of the assessment to determine the relative importance of each aspect.
- In the IB, we place the highest emphasis on creating programmes, courses and assessments that are construct relevant.
- For assessments, this means we focus on providing meaningful tasks and questions that test the higher-order thinking skills rather than examinations that are easy to mark reliably.

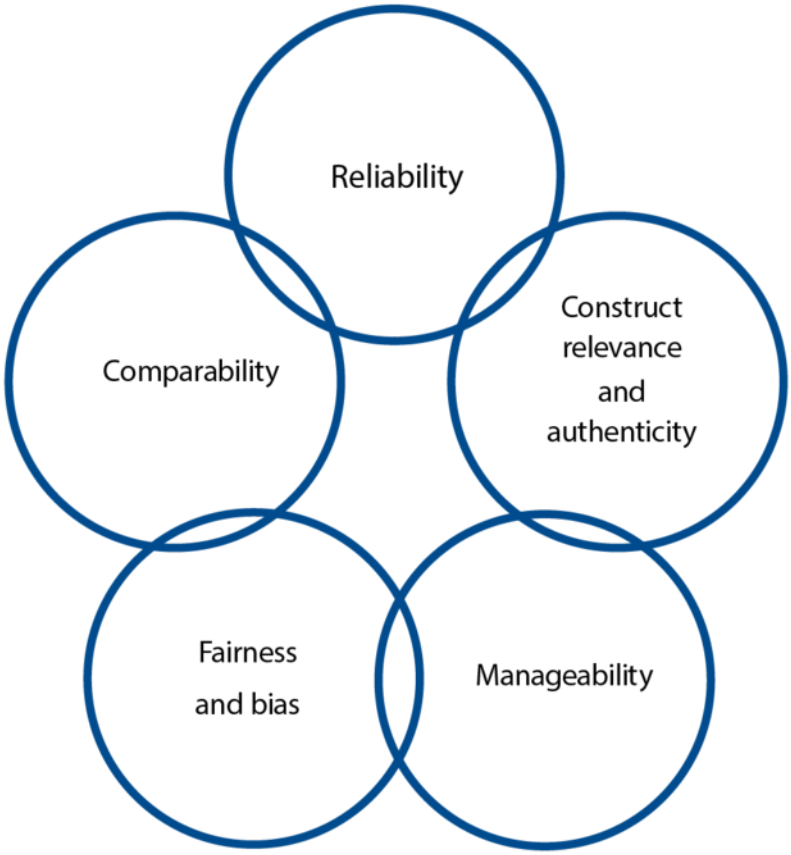
Validity (and reliability) is widely regarded as an essential characteristic of any assessment system, particularly a high-stakes one where the outcome is of great importance to the candidate or the teacher. These characteristics are in fact multi-faceted, with different types of validity and reliability.

The complex nature of validity can be expressed through the idea of the validity chain (Crooks, Kane and Cohen 1996). Each of the five elements in the chain are important in themselves but are not sufficient by themselves to make the assessment valid (that is, fit for purpose). For example, an assessment may be very reliable, but may systematically disadvantage one particular group. Alternatively, its task may focus on exactly what it intended to but, because of its duration and requirements, also test the stamina of the candidate and resourcefulness of the school.

For an assessment to be valid, all links in the chain must exist.

Figure 14

Validity chain



The sections below discuss each of these elements in more detail but, while all of the elements are necessary to achieve validity, there is also tension between each of them. For example, we want an exam to cover all aspects of the curriculum, but this can easily lead to tests that are too long for candidates. Another example might be the desire to contextualize a test for each particular culture or country (fairness) but this poses questions about whether we are really delivering the same test (comparability).

Balancing aspects of validity

In designing an assessment, it is important to balance these competing priorities, usually in the context of the purpose of the assessment. It is not possible for a single assessment to achieve the highest standards in each of these elements, so a compromise must be reached. Likewise, some of these elements are fixed during the assessment design, while others, particularly reliability and fairness, evolve during delivery of an assessment and its marking and grading.

Figure 15

Balancing competing priorities between aspects of validity: fairness versus manageability

Balance is wrong Unreasonable burden on candidates (manageability)	**Balance is right** Candidates are assessed in a reasonable way	**Balance is wrong** Too much luck is involved in what topic is tested (fairness)
Assessment contains 50 extended tasks (each 20 minutes long) covering all aspects of the course. This means over 16 hours of assessment.	Assessment contains 20 short questions (2 minutes), 5 longer tasks (10 minutes each) and 2 in-depth tasks (30 minutes). Each question is on a different aspect of the course. Total assessment lasts 2½ hours and covers just over half of the possible topics in the course.	Assessment contains one 30 minute in-depth task, testing both knowledge and understanding on one aspect of the course (out of 50). Candidate's final grade is based entirely on one task.

From the following sections it should become apparent that, despite their separate definitions, there is a considerable overlap between the way in which these five elements occur and are managed. They are also aspects of the wider concept of validity.

Finally, in the IB we place the highest priority on construct validity—that is, that our assessments test the traits and abilities that they are intended to test. However, it is important to remember that this prioritization cannot be wholly at the expense of the other aspects of the validity chain.

Reliability

Reliability is defined as "the extent to which a candidate would get the same test result if the testing procedure was repeated".

As discussed below, this is not necessarily the same as the candidate obtaining the "right result".

In their introduction to the concept of reliability (Ofqual Reliability Compendium 2011), Winkley and Cresswell (2011) highlight a (not exhaustive) list of sources of potential unreliability.

1. Inter-marker reliability. One examiner might be more or less lenient on particular questions than the next (or even the same marker might be more or less lenient from one day to the next).
2. Variability in a candidate's performance. A candidate's performance on an exam might vary a little from one day to the next, particularly if the conditions of the exam change (morning or afternoon, who is administering the test, how well they slept the night before, whether the caretaker is mowing the lawn outside, whether the candidate has a headache or not, and so on).
3. Different examination papers. Different questions will appear from one exam paper to the next which might test different facets of the candidate's understanding (tests usually sample from the curriculum because there is not enough time to test everything, and candidates may choose to revise one topic but not another).
4. Comparability of results from one year to the next.
5. Differences between examination specifications. Ensuring comparability over time can be challenging when there are changes to exam specifications and syllabuses.

6. Different types of assessment activity. Many qualifications are made up of different types of assessment activity, and these different assessment methods present different types of assessment reliability challenges.
7. Different types of questions. Candidates may perform differently depending on the type of questions they are given.

In practice, the IB usually only thinks about a candidate taking the assessment once and looks for consistency in the result of the process, and so focuses on points 1, 3, 6 and 7 from this list.

As an awarding organization, the IB takes steps to increase the levels of reliability in the assessment. Marking reliability is a central aspect of this and many of the purposes of standardization, the quality model, and moderation are to ensure that examiners mark to a consistent standard.

Reliability in marking

This quick exercise demonstrates the concept of reliability in marking and how it can work in practice.

Below are five excerpts from authentic responses to an IB language acquisition assessment.

Use the mark scheme shown below to judge how the quality of the language out of 10 and record the marks you would give to each piece of work.

After you have "marked" each piece of work, view the feedback about the quality of your marking. The feedback is shown after the candidate scripts below.

Markscheme

How effectively and accurately does the student use language?

Marks	Level descriptor
1–2	**Command of the language is generally inadequate.** A very limited range of vocabulary is used, with many basic errors. Simple sentence structures are rarely clear.
3–4	**Command of the language is limited and generally ineffective.** A limited range of vocabulary is used, with many basic errors. Simple sentence structures are sometimes clear.
5–6	**Command of the language is generally adequate, despite many inaccuracies.** A fairly limited range of vocabulary is used, with many errors. Simple sentence structures are usually clear.
7–8	**Command of the language is effective, despite some inaccuracies.** A range of vocabulary is used accurately, with some errors. Simple sentence structures are clear.
9–10	**Command of the language is good and effective.** A wide range of vocabulary is used accurately, with few significant errors. Some complex sentence structures are clear and effective.

View the scripts in turn, assigning a mark to each using the markscheme as your guide.

Figure 16

Reliability in marking exercise

Candidate A

At first, I am going to give you an idea of how dangerous websites and apps like Facebook, Twitter and Snapchat can be.
For example, you post a nice picture of you in your bathing suit at the beach and somebody you don't know immediately sends you a friend's request. Maybe the contact information is great, he or she might be in your age and lives nearby, but are you sure this is really the person that sits behind the screen and messages you at the moment?

Candidate B

Most of the time we as young people don't consider the dangers associated with loss of privacy while using social networks, and that's a huge problem, because even when the social networks protect the information it will be not ~~useful~~ helpful if the users have their profile information as "public". With this I'm not saying that "public" configuration it's bad, but it's kind of because most of all put their real information at the moment of create an account on a social network, and that makes place for problems like identity theft or even for things like cyber-bullying that is growing more every time in many countries.

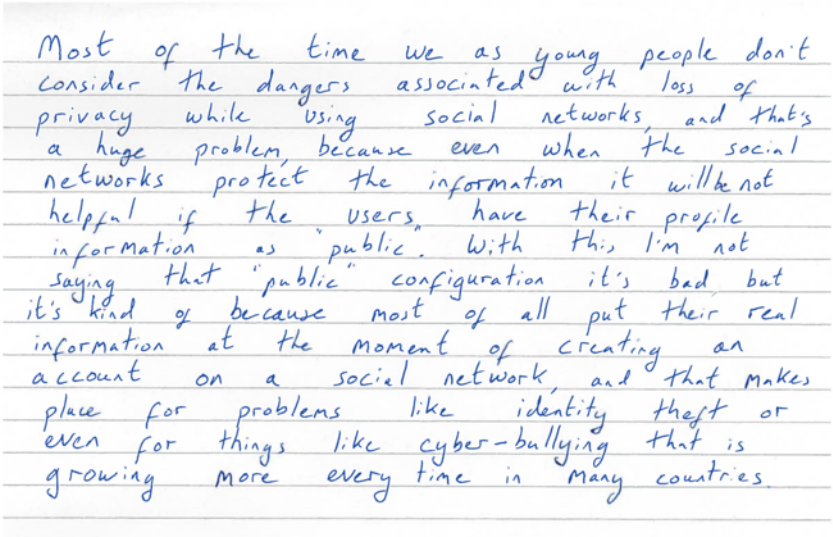
Candidate C

When people think in social networks, the first thing that comes in mind are "facebook", "Twitter" and "Instagram", am I wrong? On Facebook you can meet tons of friends from all around the world just by sending them a "friend request" and if they accept you, you are allowed to see every photo or information about this person.
So, how private do you think "FB" is? Well, actually it is not that private if you don't know to use it well. When you upload photos, you have to make sure that only your friends can see it.

Candidate D

We all use Facebook. By posting our personal information online, we could reunite our "long-lost friends", but simultaneously, other people having all sorts of intentions might be reading those information as well. Moreover, statistics revealed that social networks including Facebook make trillions of dollers every year through selling our information to companies, to individuals and even to crime groups. And the tragedy lies not only in overwhelmed advertisements, but also in potential exposure to cyber bullyings and crimes.

Candidate E



Most of the time we as young people don't consider the dangers associnted with loss of privacy while using social networks, and that's a huge problem, because even when the social networks protect the information it will be not helpful if the users have their profile information as "public". With this I'm not saying that "public" configuration it's bad but it's kind of because most of all put their real information at the moment of creating an account on a social network, and that makes place for problems like identity theft or even for things like cyber-bullying that is growing more every time in many countries.

You should now have completed your "marking" of each script and have given a mark for each candidate.

Now view the feedback on your marking.

Feedback

Did you notice that candidate B and candidate E were the same pieces of work but in different handwriting?

Did you give candidate B and E the same mark?

You may also have found yourself debating which of the two marks in a particular markband to award. How confident are you that you would make the same decision if you were to mark it again? The point we were trying to make here was not about the quality of the students' work but how the same person marking the work can make slightly different judgments each time they mark a piece of work. Broadly, your view of the quality of the work is the same each time, but the exact mark may vary slightly.

Try this exercise again by marking each piece of work again in a few days' time and see how similar your marks are a second time. (Try not to remember what mark you gave it initially.)

Consistent outcomes are not the same as the right outcome

It is important to understand that the goal of having a high level of reliability is for candidates to get the same (fair) mark whichever examiner looks at their work, not the "right" mark. How good a piece of candidate's work is relies on professional judgment and two teachers will often disagree on what mark to award. The point of reliability is that they both provide the same judgment (that of the senior examiner).

This presents a particular challenge when dealing with enquiries upon results (EUR). In this case, the examiner needs to make sure that they provide the same mark that they would have given had this been the first script they were marking, and not be swayed (positively or negatively) by any extra information they now have, such as grade boundaries or the impact on the individual candidate.

The poor understanding of assessment reliability outside of the education sector has been well documented, but with increasing public discussion in examination results this is a topic which needs greater emphasis.

> As the literature suggested would be the case, the participants found assessment reliability, and in particular measurement inaccuracy, difficult concepts to comprehend.
>
> (Chamberlain 2010: 3)

> Sometimes I think the exams aren't really a fair gauge to a person's ability, are they? They're more about how that person was at the time and how studious they were and all the rest of it.

> Somebody who's not very good at exams could excel later on and become top of their field. (Health sector employee, male).
>
> (Chamberlain 2010: 27)
>
> While about 63 per cent of teachers and students selected 'Any level of error has to be unacceptable—even just one candidate getting the wrong grade is entirely unacceptable' on one hand, over 50 per cent of them also selected 'There's a difference between an avoidable mistake —like a typo on a paper—and something inevitable like inconsistency between two markers', suggesting tolerance for error. This inconsistency may reflect the weak relationship between knowledge about reliability and attitudes to unreliability.
>
> (He, Opposs and Boyle 2010: 27)

Construct relevance and authenticity

How accurately are we measuring the thing we are trying to measure? Construct relevance is sometimes referred to as construct validity, but to avoid confusion with overall validity we will use construct relevance.

The idea of an authentic assessment is closely related to construct relevance. It means that the testing is done in a way that matches the situations in which the candidate would expect to encounter problems in the real world. Assessments that remove tasks from their context, over-simplify or are clearly contrived are examples of inauthentic assessments.

It is straightforward to create examples of poor assessment design which has a low degree of construct relevance—for example, ability to write a letter by an oral exam—but often even the accepted approach to setting tasks is not truly testing what is intended. As an example, consider how some traditional literature examinations were constructed. The candidates have been taught about the literary devices used in a set text and then the assessment will ask them to write an essay related to one of them. If the candidates simply recall everything their teacher said, are they actually demonstrating an understanding of literature?

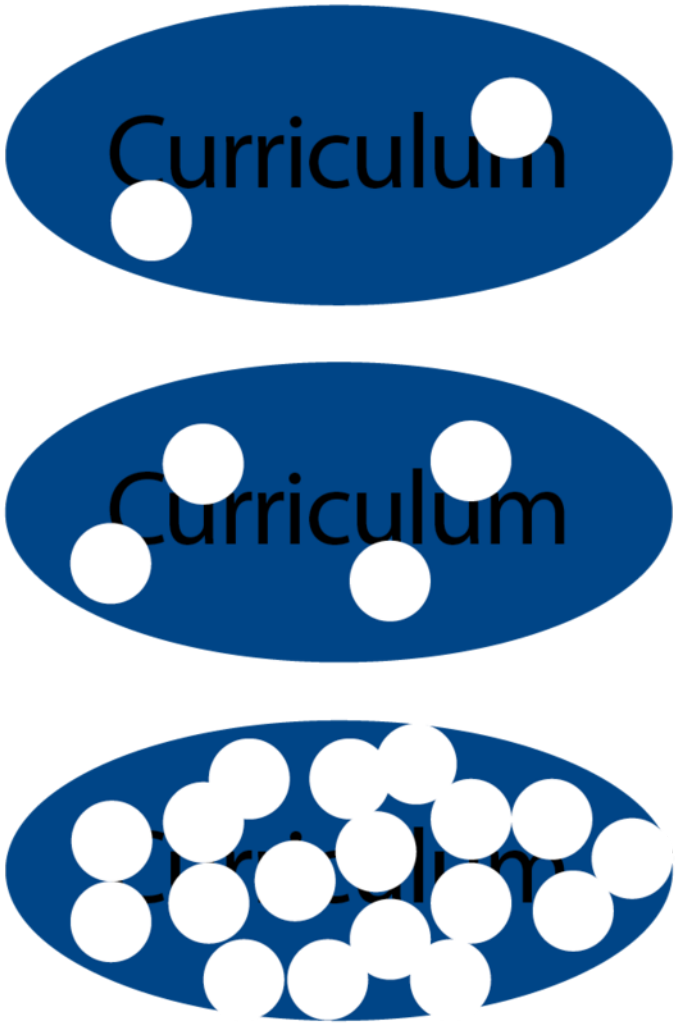
A clear understanding of what a particular assessment task is seeking to allow the candidate to demonstrate, and an inquiring and challenging review of the degree to which it does so, represent best practice in ensuring appropriate degrees of construct relevance. In particular, consider what other skills the candidate would require to undertake the task. Open tasks, such as essays or projects, are particularly vulnerable to requiring the candidate to have a high competence in writing in order to have the chance to demonstrate the research or analysis skills the task intended to address. This is encapsulated in the concept of Universal Design of Assessment which seeks to reduce barriers to any candidates where challenges are not related to what the assessment is designed to evaluate (Dolan et al 2013).

The historical approach to the measurement of construct relevance (or validity) was derived very much from a psychometric background. This led to a dangerously circular situation where what the test was assessing (known as the construct) was defined by the test itself. If all the questions in the test gave a consistent result, then this was seen as proof the test was construct relevant. However, if a question didn't align with the outcome of the other questions, it was not seen as construct relevant and so discarded. Modern approaches tend to avoid this purely statistical and self-referencing definition of construct relevance.

Construct relevance has a particularly strong relationship with the curriculum, as outlined in figure 9 called "Relationship between objectives, outcome and design". In many cases, an assessment will only ask questions based on a small selection of the material, and the question of whether this curriculum coverage is sufficient is one of construct relevance. Similarly, the choice of the types of questions to ask is an important part of this design, as different question or task types have different strengths and weaknesses in terms of the kind of construct they can test.

Figure 17

Each exam (white circle) covers part of the overall curriculum



Manageability

In contrast to the previous two topics, the aspect of manageability has not been well researched in the literature and there is no single definition or approach to its measurement. In general terms, it refers to the effort required to take the assessment, and it is reasonable to talk about the manageability of an assessment in terms of the candidate, the school and even the IB.

Candidate manageability often relates to the length of time of the assessment. An eight-hour exam is considered to be an unreasonable demand on an 18-year-old. There also needs to be consideration of the time at which the assessment takes place. For example, a series of four exams, each of an hour duration, may not be unreasonable, but if they are all on the same day with only a short break between them then this may be considered a heavy burden. Particularly for the IB, we need to consider the burden caused by other IB assessments that candidates may also be taking as part of their education—this comes back to the principle that we are considering the validity of our programmes, not just the individual subjects.

From the perspective of a school, manageability may also include the requirements to provide material for the assessment. A vocational engineering course might require each candidate to have an engine to assemble. Within the context of IB qualifications, a requirement to be able to take the assessment on a

computer or similar device would certainly demand consideration on whether this assessment was manageable.

Another aspect of school-based manageability is how the candidates' work is passed to the IB for marking. The need to record or film a presentation represents more of a demand on the school than posting a written piece of work.

Finally, manageability is also an issue for the IB in terms of the quantity of candidate work that is reviewed. Attending a three-hour drama production may provide the best evidence of a candidate's ability, but it is not practicable if the task is to be externally assessed. Another example is offering a wide range of optional questions which then require careful and time-consuming work to establish a common standard between them.

There is often a tension between manageability, reliability and construct relevance. Increasing the amount (length) of assessment will provide more evidence of the candidates' understanding of the whole curriculum and increase the probability that generous and harsh marking decisions will cancel themselves out—but it will also lead to the assessment testing the candidates' ability to maintain their performance in a long exam rather than the objectives of the assessment.

The IB places tight controls on the manageability of assessment, in particular around the total length of assessment for each course.

Fairness and bias

A test is biased if it gives an advantage to one person without that being the point of the assessment. To put this in perspective consider the following examples:

- a history exam which is written entirely in Latin
- a maths exam in which all the questions are based around scoring runs in a cricket match
- an art (painting) exam where the easel is set up two metres from the floor.

In each case, actual tasks (questions) may represent a reasonable test, but some candidates (short ones in the last of these examples) will be at a considerable disadvantage. In practice, most examples of bias are more subtle than these, but unless care is taken, bias can make a considerable difference to candidates' results. Putting questions in context is a particular challenge as situations that are familiar for some candidates will be very unfamiliar for others, especially given the international nature of the IB.

Bias can be defined as a difference in outcome of an assessment process that is not related to a genuine difference in the aptitude or achievement being measured. Bias can arise from the way in which the assessments are delivered, from the marking of an assessment (which becomes an issue of marking reliability), or from the assessment questions/tasks themselves.

Bias from the delivery of the assessment

The examples from the beginning of this section include bias in delivery, in this case the unlikely idea of an easel being too tall for some candidates, but there are many ways in which assessment, and particularly examination, delivery causes bias. The most common relates to the timing of an exam, such as during periods of extreme temperature or pollution in some parts of the world, or during periods when some students are not able to concentrate or prepare properly. This is a particular challenge for the IB as the conflicting requirements from countries around the world mean that every possible date is inappropriate for at least some schools.

Another common bias may lie in how the hall is set up. Consider two candidates, one sitting in direct sunlight and another in dark shadow. Inconsistent practices about examination rules might be another example: one candidate may be kept strictly to time while another is given some flexibility. The IB manages this through clear and consistent rules set out in the *Assessment procedures* for each programme, and by treating a breach of these as maladministration.

Bias arising from the marking

It is important to start this section by highlighting that most marking bias is unconscious rather than deliberate. The human mind is designed to use shortcuts to help decision-making, and these often result in unconscious bias. This does not mean that the IB does not have a duty to mitigate for this bias, but that there is no blame associated with it.

Bias arising in marking can occur for a number of reasons, such as personal attitude to neatness of candidate handwriting (for example, Hughes et al 1983), preferential treatment for candidate gender (where this is known or suspected by the marker), and undue attention given to factors such as formatting, punctuation and spelling, which may not be significantly relevant in some assessment contexts. Dealing with these issues is a matter of marker training and the checking of their work.

Unconscious bias based on factors such as gender, nationality or school is also well documented, and to minimize this the IB seeks to anonymize all candidate work before it is marked.

Another well-researched area of bias is known as the halo effect. In these circumstances the examiner develops a positive opinion about a candidate if their early answers are of a high quality and this results in giving them a disproportionate benefit of the doubt with later questions.

Bias related to the assessment questions

Bias does not occur if the difference relates to what you are actually trying to measure. For example, the difference in height of men and women is not a result of bias in the approach to measuring height. The assessment may not be biased, but it could still be discriminatory.

Bias arising from the assessment tasks themselves is the more significant problem of principle. In the construction of psychometric tests, any item that is shown to have unusual response characteristics during pre-testing, or which shows substantially different response characteristics for different sub-groups of the candidate population ("differential item functioning" or DIF), may be regarded as biased and removed from the test. The candidate sub-groups may be defined by gender, ethnicity, social class or language competence, in fact by any defining characteristic that could be argued to be irrelevant to the construct being tested.

However, claims of bias towards or against particular candidate sub-groups are not always self-evidently justifiable. In the early years of the development of intelligence tests, those items that gave rise to a significant difference in response between the genders came to be excluded. This was based on the understanding that there should be no difference in the construct of intelligence between males and females, and so any item that revealed such a difference must be measuring something irrelevant. Such a view is at least open to debate, and various authors have offered explanations for differences in measured intelligence between different groupings of people, relating to biological, environmental or socio-economic factors, as well as the nature of the tests themselves.

The development of intelligence testing has shown a greater concern with the reliability of measurement than with the nature of what is actually being measured, which has been moulded to suit the demands of high reliability. There may be significant aspects of a construct that are quite legitimately linked to certain characteristics of groups within the student population as a whole. The implications of such differences for teaching and learning are significant. A preferred approach is for tests to contain a balance of items that give rise to differential performance by different sub-groups of the population, so that no sub-group is disadvantaged overall. Whichever approach is adopted it can place considerable constraints on the design of a test.

When determining whether a particular question or test item is biased, care must be taken to consider how the task can be explicitly linked to the underlying construct and what the possible factors for introducing bias might be. Basing the decision on purely statistical grounds risks falling into the trap of confusing a biased assessment with one whose purpose discriminates between these groups. Goldstein (1996) and Humphreys (1986) have suggested that it is useful to distinguish between "difference", which is an objectively determined fact, and "bias", which is a judgment about the relevance of the difference. Black (1999) proposes the following six most common possibilities by which questions might be unfair in their impact on different students.

- The context in which the question is set (for example, American cultural references favour those located in the USA compared to students elsewhere).
- Essay questions on human relations favour groups in society where emotions are encouraged.
- Multiple-choice questions may favour boys.
- Coursework/project work components of assessment may favour girls.
- A question using language or conventions of one social class would favour students from that class.
- Some questions may be intelligible only within certain cultures. For example, a question about elderly people living on their own might be quite alien in some cultures, or a question involving a typical male or female role from one culture may appear very out of place in another.

In the case of gender bias, evidence of this exists, particularly in the USA, but it is not clear which aspects of the format contribute to these findings.

In designing assessments, how we should respond to such differences is not always quite so clear. Assessments should be designed so that, by means of a variety of tasks and question types, the overall impact of bias is reduced. Any form of cultural or gender stereotyping (whether explicit or not) should be avoided. The content of individual questions must be scrutinized to avoid the more obvious categories that are known to introduce unfairness, and pre-testing of questions on samples of the different sub-groups of the student population might reveal hidden cases. However, if all biased question types and possible scenarios are excluded, there is little choice left available to assessment designers and question constructors, and the resulting constraints will have a negative impact on the validity of the assessment. Apart from avoiding obvious and unnecessary pitfalls, a balanced approach to assessment design, using a variety of different types of assessment task and format, seems to offer the most reasonable solution.

There is also a concern about how many differently defined sub-groups of a population require consideration. Should account be taken of those candidates who have different kinds of learning styles, or those who are temperamentally unsuited to formal tests or examinations? As Hieronymus and Hoover (1986) have stated, if differences in interest and motivation are considered to be biasing factors, all tasks or assessment methods may be said to have a certain amount of bias. For example, passages of text used in language examinations are bound to be of more interest to some candidates than others. In the end, concern about the potential for bias in different question types and contexts often comes down to a matter of socio-politics. Equity in assessment, which includes the avoidance of bias, is a major issue, particularly in certain countries where any demonstrable bias in an assessment instrument may even lead to litigation. However, the proof of bias, as opposed to difference in performance, is often a matter of fine judgment, linked strongly to the particular social context in which the assessment is conducted.

Gipps and Murphy (1994) concluded their book entitled *A Fair Test? Assessment, Achievement and Equity* by saying "there is no such thing as a fair test nor could there be: the situation is too complex and the notion too simplistic". However, that does not mean that assessment designers and question writers should not do all in their power to reduce the impact of bias and unfairness. Gipps and Murphy also maintain the view that assessment designers should set their goal as equality of opportunity and of access to assessment, rather than the equality of outcome that is engineered by manipulating individual test items according to their response statistics. They question to what extent it would be justifiable, for example, to bring multiple-choice papers into English examinations to improve the relative performance of boys, since this would distort the validity of the assessment according to our conception of the definition of the subject.

It is widely recognized that lack of fairness in the assessment process is only one factor contributing to inequity in education, and possibly one of the less significant ones. Differential performance by different sub-groups on a test may be the result of factors quite unrelated to the test itself. There are many other sources of inequity in education that have a major impact on candidate achievement, for example, differences in the quality of teaching within a school, differences in the level of resourcing for different schools and in different geographical areas, and differences in the social circumstances and level of family support given to individual candidates. Any or all of these could significantly affect an individual candidate's prospects of educational success in a way for which no assessment process, however fair, could compensate. Smith and Tomlinson (1989), for example, found that school effectiveness was a much greater factor in determining differences in examination results than candidate ethnicity, indicating that attempts

to adjust assessment instruments to remedy differences in performance by different ethnic groups may sometimes be inappropriate.

This kind of consideration formed the rationale behind testing for aptitude rather than achievement, but it has come to be understood that assessment of pure aptitude, ability or potential, separated from social background and educational experience, is not possible. It is also not possible to regard educational achievement in an objective fashion that is independent of social context and culture. The concept of educational success is defined and measured according to the standards of a restricted section of any given society.

Removing bias for candidates with assessment access requirements

A further aspect of bias that must be countered is the potential for an assessment task to discriminate unfairly against candidates with learning support requirements such as dyslexia, attention deficit disorder or impaired vision. This is done by making sure that the conditions under which assessment tasks are taken make appropriate allowances for such candidates, so that they can demonstrate their level of educational achievement on equal terms with other candidates.

This topic is dealt with in more detail under the "Fairness for all—meeting candidates' needs" section.

Recognizing bias

It is important to keep in mind that bias can be positive as well as negative. If a task is particularly familiar to one group of candidates or easier for them to complete, this is still bias. The aim of a fair assessment is to provide an equal chance for all candidates.

We have described how bias can be introduced in the design of assessments or in the marking process. While it is very important to be proactive in thinking about bias during these parts of the assessment cycle, it is not enough just to think about potential bias, we must also look for evidence of it in candidate results, comparing how the candidate did on a particular question compared with the examination as a whole.

It is also important to base decisions around bias on evidence, not on stereotypes. Particularly with gender differences, there are a number of widely held beliefs around the type of questions that advantage boys or girls which may or may not be true for the cohort of candidates taking IB examinations. We should always use evidence from past assessments in making such judgments.

Comparability

Comparability is one of the most complex aspects of validity. Assessment outcomes are frequently used to compare candidates for selection purposes. Where two candidates have taken the same exam at the same time, we can be reasonably confident that a candidate with a grade 7 has performed better on the day than a candidate who achieved a grade 4, but more complicated comparisons are often made.

- Two candidates who achieved a grade 6 in history but answered different questions or took different options.
- Two candidates, one who achieved a grade 5 in Spanish literature in May 2014 and one who achieved a grade 5 in the same subject in November 2012.
- Two candidates, one with a grade 4 in physics the other with a grade 4 in chemistry.
- Two candidates, one with a grade 4 in mathematics the other with a grade 4 in geography.
- A European candidate with a grade 3 in computer science and a grade 6 in Chinese literature and an African candidate with a grade 5 in Japanese literature and a grade 4 in biology.
- Two fifteen-year-old candidates, one who achieved an MYP certificate and the other who took a different awarding organization's qualifications.
- Two candidates, one who achieved a grade 6 in SL Indonesian B and the other who achieved a grade 5 in HL Indonesian B.

Comparability asks whether two assessment outcomes can be considered equal in some sense. Between subjects this is particularly difficult as we are actually testing different things and then asking if they are of equivalent value.

The concept of validity being for a particular purpose rather than a characteristic of an assessment is particularly relevant here. A candidate's result in music is a better indication of their readiness to become a professional musician than the same grade in visual arts, but both results might be equivalent in predicting their readiness to study history.

The issue of comparability is made even more complex as each candidate has their own strengths and weaknesses, and so will find it easier to perform in some subjects than in others. This is a particular challenge for the IB as the cohorts of candidates taking each exam is not the same, particularly where there is a choice of courses available.

The IB seeks to maintain three principles about comparability.

1. The standard of work to achieve grades within a subject or discipline is comparable between years.
2. Grades between subjects or disciplines have a consistent meaning so that different routes to achieve the programme award (IB diploma, MYP certificate, and so on) are comparable.
3. Although the IB aims to focus on the higher-order skills, IB assessments are broadly comparable with similar exams offered by individual nations or other awarding bodies.

Measuring comparability

There are many different ways to measure how comparable two assessments are and many academic papers have been written on the topic. Coe et al (2008), in their review of the literature on inter-subject comparability, separate methods for comparing difficulties into two broad categories: statistical methods and judgmental methods.

Statistical methods focus on comparing candidates' performance on the assessments and looking for trends. This is based on the idea that, if two assessments are comparable, then, given a large enough random sample of candidates, on average their results should be the same on both.

In contrast, judgmental methods use subject experts to look at the assessment and give their considered opinion on their relative difficulty. A range of research tools and techniques are used to ensure that they are comparing like with like.

Both approaches are perceived as having serious conceptual shortcomings, and in their paper Coe et al identified six broad criticisms for each of the techniques.

Criticisms of the statistical methods	Criticisms of the judgmental methods
• Measures factors other than difficulty, such as teaching or motivation. • Multidimensionality—the subjects may not have a common trait. • Unrepresentativeness—are the statistics based on an inherently biased group of candidates? • Subgroup differences—if different subgroups of candidates get different degrees of difficulty, does this not challenge the conclusions of relative comparability? • Disagreement between methods—so can any one of them be "correct"? • Problems of forcing equality—what would be the impact on those candidates taking the qualifications?	• Breadth of criteria—to be applicable across different subjects they must be very broad which makes them imprecise. • Crediting responses to different levels of demand—examiners tend to give more credit to good answers to easy questions than weaker answers to harder questions. • Crediting different types of performance—examiners struggle to compare different types of task, for example, short answers versus essays. • Even "judgment" methods are underpinned by statistical comparisons—as they are based on the experience of how typical candidates are likely to perform.

Criticisms of the statistical methods	Criticisms of the judgmental methods
	• Interpretation and context—comparing a single terminal exam with a series of modular assessments. • Aggregating judgments—most assessments measure several criteria which must be balanced.

Proponents of either methodology are often highly critical of the alternative approach, and others argue that all the current approaches are fundamentally flawed. To research this, Coe et al applied five different ways of measuring inter-subject comparability to England's GCSE and GCE A-level qualifications. They concluded that there was a reasonably high level of agreement between the measures of inter-subject comparability, and that the differences between them were far smaller than the differences between subjects. They also found that relative subject difficulties were stable between years.

The IB maintains comparability between years/options through triangulation of examiner judgment, statistical analysis and teacher feedback (see the section on "Grade awarding (and aggregation)". We also review the comparability at subject/discipline level through both statistical (subject pairs, concurrent achievement) and expert judgment approaches.

IB's approach to validity

- The IB believes that construct-relevant and authentic assessment is more important than maximizing reliability.
- The IB believes in a rounded, holistic education. Its priority is for strong arguments of validity at programme level. We value this more highly than the validity of individual courses or optional routes within courses.

Validity is a complex and multi-faceted balancing act between a number of important and conflicting demands. There is no single right answer; where you place the balance is ultimately a judgment based on the values of the organization that is developing the assessments.

In the IB, we place a high value on testing what is important in a way that reflects the real world. The first of these points we include in the term "construct-relevant", that is, our assessments are asking what is really important for the subject, not just what is easy to mark. The other side of this coin is authenticity, which means that the tasks we set in our assessments represent meaningful tasks which reflect the meaningful way in which candidates might encounter these activities in the real world rather than being artificial and contrived.

These objectives come at a cost to other aspects of validity, most significantly, in the reliability of the assessment. Such meaningful, authentic tasks generally require a large degree of subjectivity in marking, which means accepting larger variations between examiners than if we had, for example, multiple-choice assessments. It also has an impact on the manageability of the assessment. These kinds of assessment are more challenging and time-consuming to create and to mark. They also require more commitment from candidates to understand and engage with, thus increasing candidate workload, for example, in undertaking meaningful research tasks rather than focusing just on performance in examinations.

While we accept that other groups may choose different priorities in balancing assessment validity, we are confident our position is appropriate and defensible for externally verified IB assessment.

The IB aims to do more than other curricula by developing inquiring, knowledgeable and caring young people who are motivated to succeed. It hopes its students will help to build a better world through intercultural understanding and respect. Each of the IB's programmes is committed to the development of students according to the IB learner profile.

The importance of this in assessment terms is that the purposes of the IB are defined at the programme level, not at the level of individual subjects or disciplines. Therefore, the question of validity needs to be

asked at programme level and to include the rules that govern the award of the overall certificate, not just each individual grade.

This does not mean that it is not important to consider each course, or indeed each individual assessment task, and some aspects of validity only make sense at this level of detail. However, in making any overall validity argument we need to think about the overall programme of study the student has undertaken.

Most of the principles of assessment which are outlined in this resource apply to all of an IB education, but section C looks at each programme in turn and discusses any unique feature of that study.

Defining standards

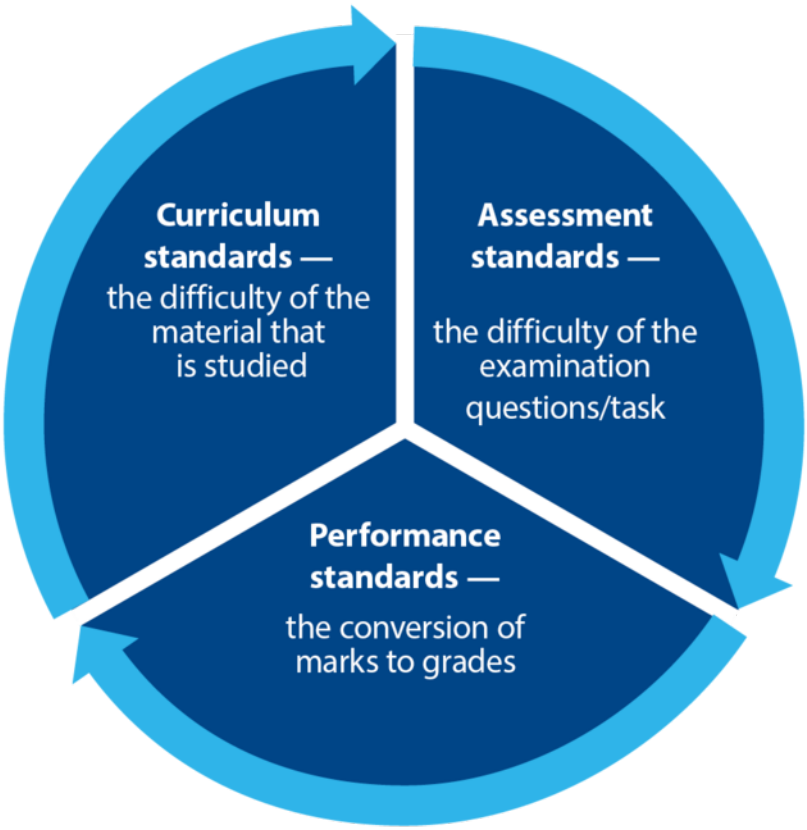
- There are three parts to any definition of standards: curriculum, assessment and performance.
- In formative assessment, the focus is usually on curriculum and assessment standards.
- The IB uses "weak criterion-referencing" which means using a balance of criterion with comparison of candidate outcomes in previous years to set the standard.
- Maintaining standards is as important as setting them.

Three meanings of standards

"Standards" in assessment generally refers to how difficult the tasks/tests we set the candidates are, and is a core part of comparability. We can generally consider the concept of a standard in three different ways.

Figure 18

The three aspects of standards



Each of these can be varied to change the overall difficulty of the subject; however, changes to each has different impacts and timescales.

For the IB, curriculum changes would generally take place during the curriculum review, although it is possible to make them outside the formal review cycle. Fairness requires that we give teachers time to

adjust their teaching to adapt to any changes. For DP and CP this would ideally be two years (the length of the course), but for MYP it could be longer given that the programme is five years in length, although typically the assessment only evaluates the last two.

Assessment standards are dependent on the paper development cycle, typically a year to 18 months. Unlike curriculum standards, assessment standards will always vary slightly from year to year as it is impossible to design two sets of questions that are of identical demand. Even repeating the same questions as the last paper will provide a lower level of challenge to candidates as, to some degree, they will be familiar to them the second time. On occasion, a decision may be made to deliberately change the assessment standards of a paper if previous papers were not performing as they needed to, and, depending on the scale of the change required, it may be possible for a paper that is already in development to be amended. Again, fairness requires that we give warning of the change to teachers, but depending on the scale of the changes this would be considerably shorter than for changes in the curriculum.

Performance standards (that is, grade boundaries) are adjusted every year to balance any change in the assessment standard. For example, if the paper is more difficult so that candidates achieved lower marks, then the grade boundaries would be moved down so that the same quality of candidate achieved the same grade as they would have done in the previous year. Unlike the other two, it is primarily examiners rather than teachers who need to understand the performance standard so that they can maintain it across sessions. It is possible to change the performance standard at the point at which grade boundaries are set, although if the IB were intending to do this it would usually warn schools in advance so they could manage and adjust candidate expectations. In such a case, the critical step is that examiners understand what this new standard looks like in order to be able to apply it in future sessions.

In setting the overall standard for an assessment it is important to balance these three different definitions. While it is possible to set extremely challenging questions on simple material, or to set extremely high performance standards on a simple set of questions, this often results in very poor levels of construct relevance. For example, requiring full marks on a test to achieve a grade 7 requires the candidate to have very high levels of accuracy. A candidate who has an excellent grasp of the subject but is slightly careless or has poor writing skills is unlikely to obtain a grade 7—is this the intended purpose of the assessment?

The definition of standard applies to both formative and summative assessments. However, in formative assessment, the setting of the performance standards is usually part of the judgment of the teacher in deciding on the feedback to provide. There is usually far more consideration of the curriculum and assessment standards to make sure that the test is appropriate for the learner and will provide useful information to inform future teaching.

Figure 19

This formative assessment is unlikely to provide any useful feedback

Norm-referencing and criterion-referencing of performance standards

- Norm-referencing means setting the performance standard on how well candidates do, for example, the performance of the top 20% of candidates.
- Criterion-referencing means setting the performance standard according to a description of what to look for in candidate performance.
- The IB uses "weak criterion-referencing" which means using a balance of criterion with comparison of candidate outcomes in previous years to set the standard.

The terms norm-referencing and criterion-referencing represent two different ways in which the performance standard in assessments can be set and maintained.

What is norm-referencing?

The technical definition of norm-referencing is often associated with standardized tests. The principle is to trial the test on a typical sample of candidates, and use the outcomes (which, by definition, should be a normal distribution or bell-shaped curve) as a reference scale by which to produce a score for any subsequent candidate taking the same test. This process of deriving a standard distribution of scores from the initial trial is called norming.

This technical definition of norm-referencing does not necessarily imply that a fixed distribution is applied to every set of test results, the fixed distribution is only used for the original norming. The distribution of scores by subsequent candidates can vary from this normal distribution.

In practice, norm-referencing is often used to refer to a process where candidates are put in a rank order according to performance and the proportions receiving each grade is fixed, for example, the top 15% would be given the top grade.

What is criterion-referencing?

Criterion-referenced assessment was first put forward by Glaser (1963). It represented a significant change in setting performance standards, putting an emphasis on measuring candidate achievement "with respect to a well-defined behavioural domain" (Popham 1978).

In criterion-referencing, candidate performance is compared against a predefined description of what is expected at each grade. This is typically done by subject or assessment experts using their professional judgment.

The limitations and difficulties of this approach are that it is very challenging to create such descriptions that are unambiguous and mean the same to all expert judges; indeed it has been argued that "no criterion, no matter how precisely phrased, admits of an unambiguous interpretation" (Wiliam 1993). The outcome of a traditional criterion-referenced test is that mastery of the relevant domain has either been shown or not shown.

In practice, both approaches have severe disadvantages. Strict norm-referencing requires strong evidence that the current test is of the same difficulty as the initial test, while criterion-referencing is subject to the Good and Cresswell effect (1988), where expert judgment does not accurately take into account the demand of the questions.

Which approach does the IB use?

The IB uses an approach known as weak criterion-referencing, which is based upon criteria but recognizes the evidence of the Good and Cresswell effect. In this approach, expert examiners are asked to establish a narrow range over which the grade boundaries could lie based on the criterion (grade descriptors) and this is then compared with boundaries calculated to match performance from previous years. Where these two boundaries align the grades are set, but if they disagree there is further discussion to establish how this contradictory evidence can be aligned.

Finally, it is important to keep in mind that criterion-referenced tests and norm-referenced tests differ more in the analysis and interpretation of candidate responses than they do in the kind of questions set.

Maintaining standards

- Once the appropriate standards have been established the IB needs to ensure they apply every year.
- Curriculum standards are reconsidered during the curriculum review.
- Assessment standards and performance standards are maintained using a mixture of professional judgment and statistical evidence.

Comparability is an essential aspect of validity. This means that we need to ensure that the same standards apply every year. It is important to recognize that IB standards are based around the meaning of grades not marks—the difference between the two is explained here.

Curriculum standards are the easiest to maintain as they are the same between sessions. Assessments need to ensure that they are a true reflection of the whole curriculum and this is monitored during paper writing. However, external factors can result in shifts in curriculum standards. The classic example is with computing skills. As computers develop and become a more familiar part of everyday life, knowledge and understanding which was once specialist and perceived as demanding becomes commonplace and routine. Therefore, the curriculum standard for the topic has changed despite the content remaining the same.

The IB has a cycle of curriculum reviews to address both this issue and to keep content up to date, and, as a result of this review, there is an expectation that the curriculum standard will change—we then need to balance the assessment and performance standards in order not to disadvantage candidates.

Assessment standards in IB examinations change every session, as no two examination papers can be identically demanding. In some education systems, extensive pre-testing is undertaken to establish the demand of each question and papers are carefully constructed to ensure a high level of confidence in the level of difficulty experienced by the candidates. In the IB, we do not undertake pre-testing of examinations,

because of the risk of questions being published before they are taken by candidates. We therefore rely on the professional judgment of our experienced paper setters and scrutineers to create consistent and balanced papers. We then adjust the performance standards based on candidates' answers to ensure the overall standard is maintained.

The two main ways of maintaining performance standards is discussed in the previous section and is the purpose of our grade award process.

Describing success—candidate achievement for summative assessment

- The focus of the IB mission statement is to develop young people who can create a better world and so, success for assessment should be where it supports this focus.
- While grades represent a very simplified view of the achievements of candidates, they allow stakeholders such as universities, employers or colleges/schools to make reasonable judgments around selection.
- If only more complex and holistic information is provided then the onus is on others to simplify that evidence to make meaningful selection decisions, and they may take less care in doing this than the IB does in setting grade boundaries.
- In our assessments we believe that professional judgment is important to achieve a meaningful outcome for the candidate, but also recognize that we must support that judgment with objective evidence to ensure we can minimize bias.
- Results need to be accurate enough not to disadvantage candidates by providing an outcome that is not a reasonable reflection of their achievement. However, the priority remains in making assessments meaningful.

The tyranny of grades—lesser of two evils

Consider the knowledge, skills and experience that go to make up an excellent chef. The knowledge—of ingredients and flavour combinations that will make up a perfectly balanced dish. The skills—in selecting and preparing ingredients, of cooking on the hob or in the oven, of presentation. And the experience—of using technique to achieve perfection, of judgment to know when something is perfectly cooked, of presenting a combination of ingredients and dishes to achieve sublime satisfaction at the table.

Figure 20

Knowledge, skills and experience

Now reduce all of that complexity into a single grade out of 7 to decide who the best chef is. The result is almost meaningless. Does it help if we increase the scale to 100, or even 1,000, grades? The answer is likely to be that it doesn't help; there may be more scope to differentiate but the fundamental issue of trying to compare different skill sets does not go away.

This is exactly the issue faced within assessment; how to represent the complexity of a learner's knowledge, understanding and ability to synthesize into a single outcome. Even if we could precisely capture all the information about the candidate it would still not mean that we could give them a grade that perfectly reflects their talents.

The alternative suggestion, which is often proposed, is not to award a grade at all, perhaps to provide each candidate with a personalized description of what they did well in and where they were less strong. Such an approach is in keeping with the principles of good teaching and learning, but has the severe drawback (which some proponents of this approach regard as a strength) that comparisons between candidates are very difficult.

To attempt to resolve this problem we return to the purpose of the IB assessment. It is to support students in being able to progress to further study or work. This means that there will need to be some kind of selection process by the receiving institutions (often universities for the DP and the CP, and schools for the MYP) to determine which students can be given these opportunities.

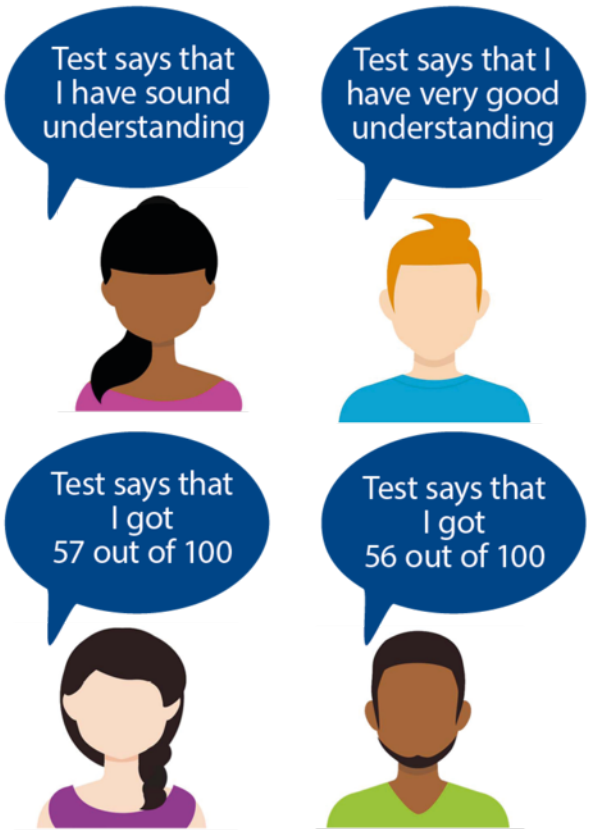
> Clearly, if the other criteria are less reliable than the examinations, greater reliance on them will lead to less reliable selection decisions.
>
> (Cresswell 1986: 37–54)

This quote featured in the introduction to this guide and it bears repeating. If selection is going to take place, then the IB has a responsibility to support its students by making it as fair and meaningful a selection decision as possible. If we only provide descriptive accounts of students to the receiving institutions, then they will need to find some way of comparing which will almost certainly be less reliable and comparable than that offered by grading examination outcomes. Consider, for example, the validity of a short interview with a tutor and all the factors that could influence the outcome which should not be the basis of selection.

This is not to suggest that summative assessments are a perfect, or even a particularly good way, of making such selection, but they are fairer than the alternatives, and most importantly the IB is constantly striving to make them as fair, meaningful and reliable a method as possible.

Figure 21

Should we differentiate between the two candidates in each pair?



In the example above, the difference between the first pair of candidates is taken from the generic grade 5 and grade 6 DP grade descriptors. If you had to make a decision between the two candidates it would probably be fair to do so, and it is likely that if they took the tests again they would get the same outcome. In the second pair, the difference is likely not to be meaningful, and the two candidates probably have the same ability.

What this example demonstrates is that not all differences are meaningful, and the use of grades provides an indication of where we feel we can differentiate between two candidates. It is certainly true that for a candidate on the boundary (either just above or just below), either grade is fair, but for most candidates a difference in grades represents a meaningful difference in performance.

This leads on to the debate of how many grades to have. If you have only two grades (pass or fail) then most candidates will get a fair grade but for those on the boundary the consequences are very serious. In contrast, if you have 20 grades far more candidates lie on a boundary, but for each candidate the consequence of being in the wrong category is less significant. This concept is explored in more detail in Cresswell (1986).

In the IB, we have generally selected seven grades as representing the number of meaningful categories our assessments can provide and also the right balance between the number of candidates on a boundary and the impact of being in the wrong grade.

Importance of professional judgment

The complex, higher-level thinking skills that form the focus of IB assessment do not lend themselves readily to simple judgment-free marking. Candidate responses are likely to be highly varied, with many

equally valid and correct forms of response. Research suggests that complex knowledge and skills should not be taught by breaking them down into small, discrete building blocks, and the same principle applies to marking such responses. When developing markschemes, they need to provide strong guidance on both marking all candidates in a consistent way and how to achieve this.

This means we need to place a great deal of emphasis on the professional judgment of markers, and particularly on the professional expertise of the senior examiners who set and explain the marking standards. This represents a strong challenge to the reliability of the assessment system, but a challenge that must be met, in the context of the precision of the outcomes.

Marking assessments

- Marking is the process of evaluating how well a candidate has completed the task they have been set.
- While the popular vision of marking may be about seeing if an answer is "correct" and scoring accordingly, this is only one possible approach. Marking can focus on the individual details of an answer or take a more holistic, global judgment.
- Other approaches to marking exist, most notably comparative judgment.
- For formative assessment, marking may not result in a numerical score but instead be purely descriptive.

What do we mean by "marking"?

Figure 22

Why marks and performance are two different things

Marking does not describe how well the candidate has done. Their level of achievement depends on a number of things such as how hard the questions were and what the expected "pass mark" is. What marking does do is to compare the candidate's answer against what a perfect answer looks like. The notion

that a camel is a horse that was designed by a committee is a good example of how inappropriate criteria (or markscheme) can produce the wrong result.

Figure 23

The perfect answer may need to be adjusted following a review of candidate work

Question—Please design a horse…

Examiner: "I think we need a new markscheme."

Marking can be undertaken in various ways depending on the nature of the task. Sometimes it is very objective—the candidate was either correct or not—and this is often the case if the answer requires only a couple of key words, or for the candidate to select from a number of possible answers.

On other occasions it is far more subjective, requiring the marker to judge whether the candidate has produced an acceptable response, or which of several statements, known as "markbands", best describes its fit to the perfect answer. Examples of markbands for DP (business management) and MYP (sport and health education) subjects are illustrated below.

Business management—Criterion B: Application

This criterion addresses the extent to which the student is able to apply the relevant business management tools, techniques and theories to the case study organization.

Marks	Level descriptor
0	The work does not reach a standard described by the descriptors below.
1	The relevant business management tools, techniques and theories are connected to the case study organization, but this connection is inappropriate or superficial.
2	The relevant business management tools, techniques and theories are appropriately connected to the case study organization, but this connection is not developed.
3	The relevant business management tools, techniques and theories are generally well applied to explain the situation and issues of the case study organization, though the explanation may lack some depth or breadth. Examples are provided.
4	The relevant business management tools, techniques and theories are well applied to explain the situation and issues of the case study organization. Examples are appropriate and illustrative.

MYP physical and health education (year 5)—Criterion A: Knowing and understanding

Figure 24

Examples of DP and MYP subjective marking

Achievement level	Level descriptor
0	The student **does not** reach a standard described by any of the descriptors below.
1–2	The student: • **states** physical and health education factual, procedural and conceptual knowledge • **applies** physical and health education knowledge to **investigate** issues and **suggest** solutions to problems set in familiar situations • **applies** physical and health terminology to communicate understanding **with limited success**.
3–4	The student: • **outlines** physical and health education factual, procedural and conceptual knowledge • **applies** physical and health education knowledge to **analyse** issues and to **solve** problems set in **familiar situations** • **applies** physical and health terminology to communicate understanding.
5–6	The student: • **identifies** physical and health education factual, procedural and conceptual knowledge • **applies** physical and health education knowledge to **analyse** issues and to **solve** problems set in **familiar and unfamiliar situations** • **applies** physical and health terminology **consistently** to communicate understanding.
7–8	The student: • **explains** physical and health education factual, procedural and conceptual knowledge • **applies** physical and health education knowledge to **analyse** complex issues and to **solve** complex problems set in **familiar and unfamiliar situations** • **applies** physical and health terminology **consistently** and effectively to communicate understanding.

Another variation is whether the marking is carried out separately for several different aspects of the work, often called criteria. For example, an essay could be measured against four separate criteria: (1) quality of grammar; (2) accuracy of key facts; (3) essay structure, and (4) quality of conclusion. Good practice with this approach is to make sure that the criteria are independent of each other. In the previous example, if a student had no conclusion he might well be penalized for this in criteria 3 and 4.

The opposite of criteria marking is global impression judgment. Here the marker internally balances all of the different aspects of an ideal answer and gives a final judgment which reflects the holistic piece of work.

Holistic versus criteria marking

There are various advantages and disadvantages to these two approaches. Holistic global impression can often be hard for two different examiners to give a consistent mark for, as they have to balance different

aspects of the work. However, the final ranked order of the work (that is, comparing different candidates' work against each other) is usually a fair reflection of which work is best.

In contrast, using criteria allows for much more consistent marking, although small differences can easily be amplified. For example, if there are three criteria, each out of 4, and a candidate is between 2 and 3 in each, then two reasonable examiners might give that candidate 6 (two in each criteria), 9 (three in each) or any score between—this is quite serious as the total possible mark is only 12.

The other challenge with criteria marking is that it can sometimes produce results that are correct but do not feel fair, if one candidate has been very good at matching their work to the criteria compared to another whose work is "good" but does not match well with the criteria.

Alternative forms of marking

Most people have an image of marking a piece of work that resembles the following picture.

Figure 25

Typical view of marking

However, this is not the only approach that can be taken to marking; one which has been particularly well researched in recent years is comparative judgment.

The basis of comparative judgment (CJ) is that the human mind is better at making comparisons between two objects than it is at making judgments against an abstract scale. As an example consider the first image below. On a scale of 1 to 10 for heat (where 10 is extremely hot) what score would you give it? Now compare with the second image. Which of the two images is the hottest?

Figure 26

Which of the two is the hottest—coffee or volcano?

While this example is very simplistic, the same argument does hold true with subjective decisions such as how well a candidate's essay has answered the examination question, compared with which of two essays has answered the question better.

The approach behind comparative judgment is that examiners make many of these "better or worse" (usually termed "win/lose") decisions and these are combined together to create a ranked order of the items being assessed. Using mathematics, it is possible not only to deal with unexpected cases (A wins against B, B wins against C but unexpectedly a judge decides C wins against A) but also to establish how big a gap there should be between two items in an order by calculating how likely each possible outcome (win/lose) is. This can then be used to match to what looks like a traditional mark.

The most important aspect of using CJ with several judges is that they must all agree what makes a "good" answer. In many situations, such as the heat example above, this is obvious, but when using it to mark candidates' work it is important to be explicit in this. For example, a history essay is "good" if it makes a convincing argument supported by accurate information, rather than being good if it is very long with lots of unconnected facts. This description of what "good" looks like is known as an importance statement.

The second aspect of CJ is that the final "marks" are based on the judgments of every examiner who looked at a particular essay as part of the process, not just one examiner. This means that the marks tend towards being a consensus of the views of all the examiners, not just one. This is very different to the usual approach employed in assessment and requires a new interpretation of what "reliability" means.

The most significant disadvantage of the method is that it requires many more marking decisions. Rather than each item being looked at once, they are looked at several times, albeit usually for less time. This does make it time-consuming to mark using CJ and to counter this, a variant known as adaptive comparative judgment (ACJ) has been developed which focuses on creating a consistent result from fewer win/lose decisions by focusing on where judgments are most needed.

CJ is most likely to provide benefits to the IB where we have highly authentic and meaningful tasks which are challenging to mark reliably. For further information on CJ and ACJ please refer to the "Bibliography".

Marking and formative assessment

Marking does not need to be simply numeric. It is perfectly possible to compare a piece of work to the "perfect" answer and provide descriptive comments on the similarities and differences between the two. This makes it very difficult to compare how good two answers are with each other, but for formative assessment, where the aim is to provide feedback to support learning, this may well not be necessary.

What is a good assessment?

- There is no single answer to this question. It depends on the relative importance placed on different priorities, and the purpose of the assessment.
- At the IB, the underlying principle is to test what is important rather than judge as important what we can test.
- IB assessment seeks to have a positive backwash effect on teaching and learning.
- In general, assessments should include a range of tasks and include the opportunity for more in-depth classroom-based activities as well as examinations.
- While there are lots of technical details about what validity for an assessment means, at its heart the definition should be that it is a good evaluation of the goals of the assessment.

This simple question is actually very difficult to answer. The reason is that different people will have different priorities and so there are no right or wrong answers.

For the IB, the underlying principle is to **test what is important** rather than judge as important what we can test. This needs to be balanced against all the other considerations such as reliability and candidate workload.

It is important to realize that it is difficult for any single approach to be successful in delivering every possible priority. In particular, good assessment design is different for summative and formative assessment. Expanding on this principle, the IB's views on what makes good assessment can be summarized as:

- supporting curricular goals
- using a range of assessment tasks
- considering wider student competencies and higher-order thinking skills.

We will now consider each of these in turn.

Good assessment supports curricular goals

- Assessments should encourage good teaching (positive backwash)
- Predictability in assessments

Assessment should not be considered as separate to teaching and learning. IB assessment outcomes are based on summative assessment and are not intended to provide direct feedback on teaching and learning. However, it is well understood that what is included in the assessment will have an impact on what is taught. This is known as the backwash effect.

The IB's principle is that the design of assessments should encourage the most desirable educational outcomes for students. The impact on student learning remains an essential consideration in the design of our assessments and, together with construct relevance (that is, testing the right thing), are our priorities in deciding how we balance the different elements of validity.

The strong impact of high-stakes assessment on teaching and learning can be used to advantage by designing assessment instruments that encourage good pedagogy and constructive student involvement in their own learning, while taking account of recent thinking in learning theory (for example, Murphy (1999)).

The desired personal characteristics of students, expressed in the IB mission statement, fit very well with a constructivist theory of student learning, in which students actively engage in the learning process, take

responsibility for their own learning, and enlarge their knowledge, understanding and skills through inquiry.

Sympathy with cultural perspectives other than the student's own is expected in the assessment requirements of a number of subjects. The more affective qualities of caring and compassion are more difficult to include in formal assessment, but nevertheless must be represented within the overall assessment system. This is partially achieved through elements of the curriculum which are not assessed, such as the community service element of the MYP, and the creativity, activity, service (CAS) requirement in the DP and service learning in the CP. However, ethical working practices and understanding and valuing differences are also captured in IB assessments.

In terms of the design of our courses, the IB places a strong emphasis on predictive validity (the degree to which the results predict future success), with an awareness that the manner in which assessment is conducted will have a major impact on how IB courses are taught within schools. The assessment model (collection of assessment instruments) applied to each subject is designed to be broadly based, including a variety of types of evidence, to support construct relevance by giving the broadest range of evidence to support student achievement and learning.

While the IB is very aware that assessments have a backwash effect on teaching and learning, we also encourage schools to adopt pedagogies which develop all the goals and philosophies of the IB programmes in the students.

The IB regularly undertakes research studies to evaluate the extent to which we have been successful in designing programmes that are good preparation for further study. For more details on the range of research undertaken by the IB refer to the "Research" section of the IB website.

What is good predictability?

Predictability is the state of being able to gauge what and/or when something will happen. In assessment, this means the ability of schools to determine what questions will be asked on a paper, and when. Good predictability is essential for IB working practices in assessment as, by adhering to it, it means the IB remains loyal to the requirements of their constructs, as published for teachers, leading to a "fair" assessment opportunity in terms of curriculum alignment. What the IB has said would be assessed, will be assessed.

The IB seeks to ensure that schools' investment in their teaching options are rewarded over the whole lifetime of a particular curriculum (before it is reviewed). The entire syllabus should be examined in a way that the specific assessment dictates. Care is taken to eliminate the inevitability towards the end of a course of bad predictability (where a school identifies what has not yet been asked, and therefore is likely to appear on a paper).

The underlying principle is that nothing should be a surprise, either for the candidate or the school. The questions asked should be explicitly supported by subject guides for any given component. Where possible and appropriate in assessments, the IB seeks to reduce the likelihood of problems caused by predictability by assessing skills in the context of assessments that are designed so that candidates with pre-prepared answers are not advantaged (for example, the balance between knowledge, understanding and engagement with an unseen stimulus).

Assessment design is paramount in ensuring that there are sufficient ways to test any given theme, option, or text, to mitigate bad predictability.

It is acceptable that themes that are loosely associated can be tested across papers, where the assessment model allows for this. This is very much in keeping with the IB way of being educated which is anti-isolationist when considering any subject or option.

Figure 27

What is good and bad predictability?

Good predictability	Bad predictability
If every permutation of question was mapped based on theme/option/text plus command term, the	Teachers overly prepare candidates on certain questions, where they have noticed a tendency for a very similar question to be asked session on session.

Good predictability	Bad predictability
questions asked by the IB for that subject would feature in the results.	This is problematic for longer-response questions that require multiple skill demonstration (knowledge, analysis and evaluation).
An unpredictable approach to reusing questions.	Limitation in questions that can be asked based on ineffective assessment design decisions (for example, setting a prescribed text for a course that has only one main theme on which questions can be asked). This leads to bad predictability by design.

Good assessment uses a range of assessment tasks

A multiple-choice question, a short-response question, an extended-response question, an essay, a project, a single piece of work from a portfolio, and a research assignment are all examples of the range of assessment tasks.

An assessment instrument or component is made up of one or more tasks that are collected together, for the sake of thematic or content continuity, or for convenience. An examination paper, portfolio of work, project or research assignment are examples of assessment instruments, or components. There is overlap between the concepts of an assessment task and a component. Sometimes, a candidate may carry out only one task out of a number of choices available for a component.

There are a number of reasons why a wide variety of types of assessment task and component are used in the IB. First, from a historical and pragmatic perspective, Peterson (2003) says of the original development of DP assessment that "we had both an obligation and an opportunity to take into account the differing techniques of assessment used in those countries to whose institutions IB candidates were mostly seeking entry" and this principle extends to the CP and MYP. There are also validity considerations, relating to fitness for purpose, that require a varied approach to assessment. Finally, a variety of assessment techniques helps to reduce the potential for inequity in assessment, see also Linn (1992); and Brown (2002). The range of components and the setting of tasks within them ensure that, taken across the assessment model for a whole subject, candidate achievement is adequately represented against all the objectives for that subject.

The role of classroom-based assessment and internal assessment

Classroom-based assessment offers a number of opportunities to test candidates in areas that are not well suited to examinations. The most important aspect of this is that candidates can be asked to perform an extended task that gives them the opportunity to investigate a problem and show how they develop their thinking without the time pressures that are inherent in an examination. This means that there are a wide range of assessment tasks that can only be delivered using classroom-based assessment.

Examples of the kind of tasks that lend themselves to classroom-based assessment tasks include project work, fieldwork, laboratory practical work and mathematical investigations. Oral examinations which require a teacher to ask questions and respond to the candidate also requires administration in a classroom context, although developments in on-screen technology may change this in the future.

There are other advantages to internally assessed work within the context of an international qualification. Such work can be very flexible in the choice of topic, while continuing to address a common set of skills. This allows schools to place study in a local, cultural or geographical context, or to draw closer links between the classroom and the world outside. International schools, whose students often have a different cultural background from that in which the school is embedded, can use internally assessed work to develop a closer involvement in the local society or environment. Alternatively, internal assessment can be used in a different fashion to develop links with distant cultures, generally by contact with schools in other parts of the world. Brown (2002) also points out the value of internal assessment in allowing for cultural diversity within DP assessment. This encourages a "broader perspective of internationalism", both by

allowing for a multiplicity of cultural approaches and by giving individual students the opportunity to experience a range of cultural values.

Additionally, internal assessment can often provide individual students with the opportunity to select their own topic or issue, following a particular interest and giving students greater control over their own learning. This flexibility of approach makes internal assessment a valuable addition to students' education, improving the validity not only of the assessment process, but also of the learning experience as a whole.

There are also challenges around classroom-based assessment. One of the two biggest is that it is much harder to ensure that the candidate is not engaging in academic misconduct, for example, getting someone else to produce their work for them. With the availability of the internet this becomes even more challenging. The school is best placed to identify where candidates are not submitting their own work and more information about the resources that are available to support school leaders can be found on the IB website.

The second of these challenges is that classroom-based assessment can generate a significant burden on both the teacher and candidate. Undertaking assessment in class reduces the amount of teaching time available, and internally set tasks are usually substantial and require a significant time commitment from the candidate. While it is appropriate for teachers to spend a considerable amount of time preparing candidates with the skills and processes required for internal assessment, there may be a strong temptation, felt by both candidate and teacher, to rehearse and practise the particular task set for internal assessment more than necessary, to make it as good as possible, further reducing the time spent teaching.

Classroom-based assessment can be assessed either externally or internally. With external assessment the work produced by the candidate is sent to the IB and assessed by an examiner. The quality of the examiner's decision can then be monitored in our usual way (see section on "Marking"). Classroom-based assessment can result in long pieces of candidate work which are challenging to mark.

Internal assessment means that it is the school (typically the candidate's teacher) that marks the candidate's work, and the IB then checks that the teacher has correctly applied the global standard through a process of moderation.

There are different views of internal assessment around the world. Some systems place great value on the fact that teachers are best placed to give a holistic opinion on the performance of the candidate rather than just having a "snapshot" from a single assessment. Other education systems rate teacher performance based on their candidates' outcomes, which creates a strong incentive for teachers to award high marks in any internal assessment.

A teacher's judgment can also be affected by past experience of a candidate's work, which establishes certain expectations. Teachers may sometimes be unclear about the limits of their role in guiding and supporting candidates as they carry out internally assessed work, and may often have only a limited view of global standards of achievement within their subject area. When assessing their own candidates' work, teachers may be heavily influenced by the general standards existing within their own schools. Even where there are no incentives to over-reward candidate work, the professional relationship which teachers establish with learners makes objective decisions challenging. Research on this unconscious bias suggests that it can have a noticeable effect.

The IB believes that the benefits gained from being able to set meaningful assessment tasks that can only be assessed through classroom-based work which is evaluated by someone who has seen the candidates' development are greater than the risks internal assessment creates. Therefore, an internal assessment task will usually be part of any set of assessments.

Finally, classroom-based tasks may also provide the opportunity for candidates to demonstrate aspects of an IB education that are not assessed.

Collaborative working versus individual marks

One challenging area is how to assess collaborative tasks, where we expect candidates to work together in completing an activity. This is an important aspect that candidates will encounter in their future workplace but one which is often neglected in academic assessment.

In some circumstances, it is possible to identify the individual's role in the group activity. An example might be a dance performance where we can see the skills demonstrated by each individual dancer and so mark them individually. In these situations, group work is not contentious.

In other situations, it is not possible to identify who is responsible for what aspect of the final work. In these situations, it is possible to award a single mark for the entire group, but this does not take into account differences in the contribution each candidate has made to the overall result. One candidate might have contributed most of the work and expertise, but would get no more credit than any other candidate. Given the individual nature of the way that IB results are used to make selection decisions, in general we do not think this is a fair approach to take and so generally try to avoid group assessments where individual candidate achievement cannot be measured.

This approach can create cultural bias as the idea of individualism is traditionally a western European ideology. The IB recognizes this issue but sets it in the context that even in alternative cultures the assessment outcomes are often used to determine individual selection decisions.

Good assessment considers the wider student competencies and higher-order thinking skills

An IB education seeks to achieve more than getting students to learn "facts". This long-standing goal of the IB is reflected by current thinking among governments for the need to provide students with 21st century skills, workplace competencies or similar initiatives.

The IB's philosophy and approach to student competencies is focused on the learner profile and its link to international-mindedness. This will be explained in more detail in the section on continuity between programmes.

A good assessment considers the full range of outcomes that the course seeks to achieve and allows the candidates to demonstrate their abilities in all of these. However, it is often the case that it is only desirable or manageable to measure a small fraction of these outcomes, and good quality assessment also balances these limitations. For the IB, these outcomes can best be categorized by higher-order thinking skills, wider student competencies and their link to international-mindedness.

Higher-order cognitive skills

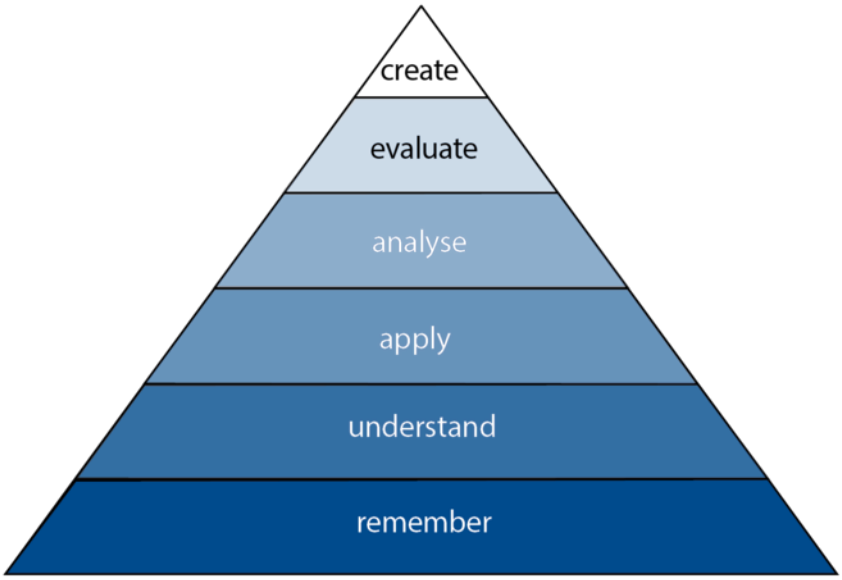
IB assessments seek to test the higher-order thinking skills of evaluation and analysis, not simply knowledge recall.

This point was made by Peterson (2003), whose views shaped the educational philosophy of the IB. He stated that "what matters is not the absorption and regurgitation either of facts or of pre-digested interpretations of facts, but the development of powers of the mind or ways of thinking which can be applied to new situations and new presentations of facts as they arise". Sugata Mitra, addressing the IB Heads Conference in 2011 in Singapore, took the argument a stage further, arguing that the immediate and comprehensive availability of knowledge through the internet means that knowledge itself has little value —it is the ability to analyse, interpret and select knowledge that is required by 21st century citizens (Mitra 2011).

IB assessments have long attempted to give significant attention to these so-called "higher-order" cognitive skills (Bloom et a 1956; Anderson, Krathwohl 2001). There may be disagreement about the hierarchical nature of the levels Bloom proposed, or about the number of levels, but his taxonomy of educational objectives still provides a useful framework through which to express the diversity of skills required. Bloom's higher-order skills certainly require the use of a different kind of assessment. Student skills of analysis, synthesis and evaluation can only be properly gauged by requiring them to analyse, synthesize and evaluate at some length. Performance assessment is the only realistic means of measuring student achievement in these areas, and because the outcomes of such activity cannot be tightly prescribed, these assessments must be relatively unstructured and open-ended where there are many diverse but correct responses.

Figure 28

One possible way of describing a hierarchy of thinking skills is Bloom's Taxonomy



The ability of our assessments to recognize and reward a student's performance in these skills is essential if they are going to be meaningful, despite the challenges this presents to reliability and other aspects of validity. Tests which only reward the recall of knowledge, concepts and routine techniques are not fit for purpose within the goals of IB education.

Student competencies and the learner profile

> Education today is much more about ways of thinking which involve creative and critical approaches to problem-solving and decision-making. It is also about ways of working, including communication and collaboration, as well as the tools they require, such as the capacity to recognize and exploit the potential of new technologies, or indeed, to avert their risks. And last but not least, education is about the capacity to live in a multi-faceted world as an active and engaged citizen. These citizens influence what they want to learn and how they want to learn it, and it is this that shapes the role of educators.
>
> (Andreas Schleicher 2016)

It is increasingly being claimed that the skills that are required this century are fundamentally different to those of previous generations. While there are those who would argue that the inquiry approach that underpins these 21st century skills has been valued since Socrates, there is general agreement of the importance to provide students with a wide range of attributes to prepare them for life. See, for example, the arguments made in Llewellyn (2014).

There is a wider range of different ways of categorizing these skills including OECD's 21st century competencies, RAND Education, NRC Framework and others. Within the IB, we describe these competencies within the learner profile.

Figure 29

IB learner profile

Not all aspects of the learner profile are appropriate to measure through summative assessment, but several are encapsulated within the concept of higher-order thinking skills. Good assessment recognizes the importance of these characteristics and even when it is not designed to measure them, it can offer students a chance to develop these competencies. Examples of this could be through encouraging ethical (principled) approaches to surveys and experimentation, supporting appropriate peer review and introducing unexpected contexts to students.

For more details on the IB's wider approach to student competencies refer to the material available on the IB website or the relevant programme's *From principles into practice* document.

International-mindedness and intercultural understanding

IB programmes are studied by students in many countries and of many nationalities. As well as the academic aims of our programmes, the IB intends that students should develop as "caring young people who help to create a better and more peaceful world through intercultural understanding and respect", and "who understand that other people, with their differences, can also be right" (IB mission statement 2002). There is, therefore, both an international context and an intercultural understanding purpose to IB teaching, both of which must be reflected in the assessment.

The most important step in delivering this is through having academic experts, including examination paper authors and curriculum developers, from a wide range of cultural backgrounds. It is important to the IB mission not to obscure differences but to engage with them in a way that allows students to explore them without being disadvantaged.

In some subject areas, the issue of cultural variety can be encouraged through a recognition of different cultural emphasis in the curriculum. Examples of this approach can be found in biology, chemistry,

psychology and visual arts. In the first three of these, the option structures within each subject allow schools to select course content that will, to a certain extent, suit particular cultural traditions of teaching the subject.

In other subject areas, international-mindedness is encouraged through the material and inspiration the student is encouraged to use. Examples of this includes the arts subjects, literature and language but it also can be included through a wider range of internal assessment tasks.

There is more to international-mindedness than just knowledge and understanding of other cultures. Attitude and action are also important attributes. Attitudes are difficult to assess through normal school assessment, which focuses on achievement rather than affective attributes.

Within the IB programmes, this is addressed through the non-assessed elements of the course such as the creativity, activity, service (CAS) part of the DP and the community project in MYP. As the diploma cannot be awarded without candidates having completed this aspect, these non-assessed elements have a significant impact on the overall outcome of IB assessment.

Figure 30

Principled action

While allowing candidates to choose which questions to answer might be seen as the best way of addressing the different international requirements in assessment, this then poses assessment problems in terms of maintaining comparability across the options. This always occurs when there are choices of question, or very open-ended assessment tasks. It is challenging to even define what "equal demand" means when the candidates come from very different educational backgrounds. In general, it is easier to maintain comparability by setting common tasks which allow candidates to introduce their own experiences into the answers. In such cases, the challenge falls upon the examiner to maintain a common standard, but this is one step easier than having two separate tasks of potentially different levels of demand which must then also be marked to the same standard.

Information on comparability can be gained through analysis of candidate performance and this analysis is discussed further in the section on "Grade awarding (and aggregation)".

Assessment carried out in an international context has additional challenges in terms of equity, above those normally encountered within a national system. Questions that might be perfectly appropriate in one national setting become inappropriate in another. Questions referring to sports, travel, entertainment, historical events, even the weather, must be prepared very carefully. It might seem that the only way around this problem is to prepare examination questions that are devoid of all but a lowest common denominator of sociocultural context. However, to do so would not only make examination questions very limited and dull, it would also be against the whole philosophy of IB assessment and against good assessment practice in terms of ensuring validity through context-based tasks. Contextualized work and assessment are vital to good learning.

There are two possible ways around this dilemma. First, background contextual information can be provided to candidates, through specification in the subject syllabus content, by providing case studies on which questions are based, or even in the examination question itself (as long as this is not too lengthy and thus distracting from the purpose of the assessment).

A second method is to use more open-ended assessment questions and tasks that allow candidates to select their own context in which to respond. In the latter approach, the focus of marking must be on deeper levels of understanding, rather than on straightforward knowledge of subject content, since there will be no common basis of content. This is very much in keeping with the IB assessment philosophy.

Figure 31

Range of cultural norms/contexts

Even with the application of both these methods, candidates may find themselves dealing with assessment tasks having contexts that are not familiar to them within their own sociocultural background. This again is in keeping with our assessment philosophy, in that one of the aims of the programmes is to make students more open-minded to other ways of doing things, more globally aware, and more competent at operating in a non-familiar cultural environment. Part of the requirement for higher-order thinking is that students should be able to apply knowledge in unfamiliar situations. It is quite appropriate for such elements to be included in assessment, as long as they affect students from different cultural backgrounds evenly.

A significant proportion of IB students enter for examinations in a language that is not their best. Nearly all such cases relate to English, because students working in French or Spanish (the other two main languages in which IB assessment is conducted) tend to be native speakers. Considerable extra care has to be taken in the wording of questions so as not to disadvantage second-language speakers. This is dealt with in paper editing.

Our summative assessment, along with the great majority of formal assessment systems, is highly individualistic. As pointed out by Brown (2002), this is largely because the DP falls within the western European tradition, and western European societies are individualistic in nature. Candidates are assessed almost exclusively on what they achieve on their own. This may be said to be culturally inequitable, since there are a number of cultures in which the contribution of the individual is always subservient to that of a larger group; it is what the group achieves that matters. It is also the case that, in terms of individual equity, there are some people who work better in a team than they do individually, and vice versa. Additionally, it is common practice, both in the classroom and in the world of work, for individuals to work interdependently rather than independently.

What does good on-screen assessment look like?

Earlier, we set out how on-screen assessment can offer better quality assessments by using the computer as the means to ask questions that would be impossible in a paper examination. This leads to three clear points that define what good on-screen assessments look like.

- On-screen assessment uses the opportunities for the computer to set more valid tasks, whether because of authenticity and relevance, reduced bias or any other factor.
- It removes barriers to candidates engaging with the assessment or improves the overall process, for example, by removing risks and delays around posting.
- It does not create any new barriers to candidates completing the assessment, especially clumsy interfaces or technical problems.

In contrast, if the on-screen assessment uses features (such as video material) simply because they are available rather than to enhance the assessment validity then this is an indication of poor quality on-screen assessment. Technology can, and should, support the assessment process but should not drive it.

IB's principles of assessment

The following five points summarize the underlying principles of IB assessment.

IB assessments must:

1. be valid for the purposes for which they are intended. This means they must be balanced between the conflicting demands of construct relevance, reliability, fairness (that is, no bias), comparability with alternatives and manageability for candidates, schools and the IB
2. have a positive backwash effect, that is, their design must encourage good quality teaching and learning
3. be appropriate to the widest possible range of candidates, allowing them to demonstrate their personal level of achievement
4. be part of the context of a wider IB programme, not considered in isolation. Does it support concurrency of learning and the overall learner experience?
5. support the IB's wider mission and student competencies, especially inquirers, knowledgeable, thinkers, communicators and internationally minded.

What do we mean by a practice?

- These practices cover the summative assessments in the DP, CP and MYP.
- They do not apply to the PYP where the IB does not provide summative assessment, or any formative assessment the school may undertake.

This section outlines the practices by which the IB produces candidate outcomes for those candidates who enter our externally marked or moderated assessments. It does not cover any other assessment which is not conducted by the IB, for example the teacher-constructed assessment which may be part of a PYP programme.

A principle sets out **why** we do something, and a practice describes **how** we do it. So in this section we will explain the high-level practices we use to make sure our assessment outcomes are valid.

The next level of detail is our procedures which describe the individual steps in delivering each practice. Where these processes relate to schools, details can be found in the programme-related *Assessment procedures*.

Reporting candidate achievement

- The focus of the IB mission statement is to produce young people who can create a better world, and so success for assessment should be where assessment supports this aim.
- The IB grades have meaning, and grade boundaries are set taking this meaning into account.
- While grades represent a very simplified view of the achievements of candidates, they also allow stakeholders—such as universities, employers or colleges and schools—to make reasonable judgments around selection.
- If only more complex and holistic information is provided, then the onus is on others to simplify that evidence to make meaningful selection decisions, and they may take less care in doing this than the IB does in setting grade boundaries.
- Candidate achievement is more than just examination results, and even when dealing with grades remember that what may be a disappointing result for one candidate will be a great achievement for another.

The International Baccalaureate aims to develop inquiring, knowledgeable and caring young people who help to create a better and more peaceful world through intercultural understanding and respect.

(IB mission statement 2002)

What do IB grades mean?

The outcomes of a candidate taking IB assessments are grades. These grades describe the standard of work which the candidate has shown in their answers.

The IB publishes descriptions of each grade. These descriptions are different for DP, CP and MYP as they reflect the standard of work we expect from different aged candidates.

Figure 32

Examples of Diploma Programme grade descriptors

Group 3 (individuals and societies) grade descriptors

Grade 7

Demonstrates conceptual awareness, insight, and knowledge and understanding which are evident in the skills of critical thinking; a high level of ability to provide answers which are fully developed, structured in a logical and coherent manner and illustrated with appropriate examples; a precise use of terminology which is specific to the subject; familiarity with the literature of the subject; the ability to analyse and evaluate evidence and to synthesize knowledge and concepts; awareness of alternative points of view and subjective and ideological biases, and the ability to come to reasonable, albeit tentative, conclusions; consistent evidence of critical reflective thinking; a high level of proficiency in analysing and evaluating data or problem solving.

Grade 6

Demonstrates detailed knowledge and understanding; answers which are coherent, logically structured and well developed; consistent use of appropriate terminology; an ability to analyse, evaluate and synthesize knowledge and concepts; knowledge of relevant research, theories and issues, and awareness of different perspectives and contexts from which these have been developed; consistent evidence of critical thinking; an ability to analyse and evaluate data or to solve problems competently.

Grade 5

Demonstrates a sound knowledge and understanding of the subject using subject-specific terminology; answers which are logically structured and coherent but not fully developed; an ability to provide competent answers with some attempt to integrate knowledge and concepts; a tendency to be more descriptive than evaluative although some ability is demonstrated to present and develop contrasting points of view; some evidence of critical thinking; an ability to analyse and evaluate data or to solve problems.

While these generic grade descriptors should be the same for all subjects in a programme, the IB often puts them in subject-specific contexts to make it easier to understand what it means in each case. It is important to understand that the standard is not changed by this subject context. A grade 4 in a language should mean the same thing as a grade 4 in a science or a grade 4 in an arts subject. This is inherent in the IB approach to programmes where all grades count equally, but there has been much discussion among educationalists on whether such a concept makes sense—how can you compare achievement in two different subjects? Is it even meaningful to try? This concept is explored more fully in the section on "Comparability".

Figure 33

How can you compare these two pieces of work?

9. (a) $x = e^{3y+1}$

taking the natural logarithm of both sides

$$\left(f^{-1}(x) \right) = \frac{1}{3} (\ln x - 1)$$

(b) coordinates of Q are (1,0)

$$\frac{dy}{dx} = \frac{1}{x}$$

at Q, $\frac{dy}{dx}$

$y=x-1$

(c) let the required area be A

$A=\int_1^e 1dx-\int_1^e \ln x dx$

use integration by parts to find $\int \ln x dx$

$$= \left[\frac{x^2}{2} - x \right]_1^e - \left[x\ln x - x \right]_1^e$$

$$= \frac{e^2}{2} - e - \frac{1}{2} \left(\frac{e^2 - 2e - 1}{2} \right)$$

In the context of IB assessment, the argument comes back to the validity of purposes of our grades. They are intended to allow stakeholders to compare students' attainment, and therefore it is meaningful to use statistical and qualitative methods to aim for parity in the meaning of grades.

What is the difference between marks and grades?

Marks and grades are not the same thing.

Figure 34

It is not just how far you walked, but also where you are walking

We would expect "good" candidates to have completed most of the task.

We would expect "good" candidates to have only completed part of the task.

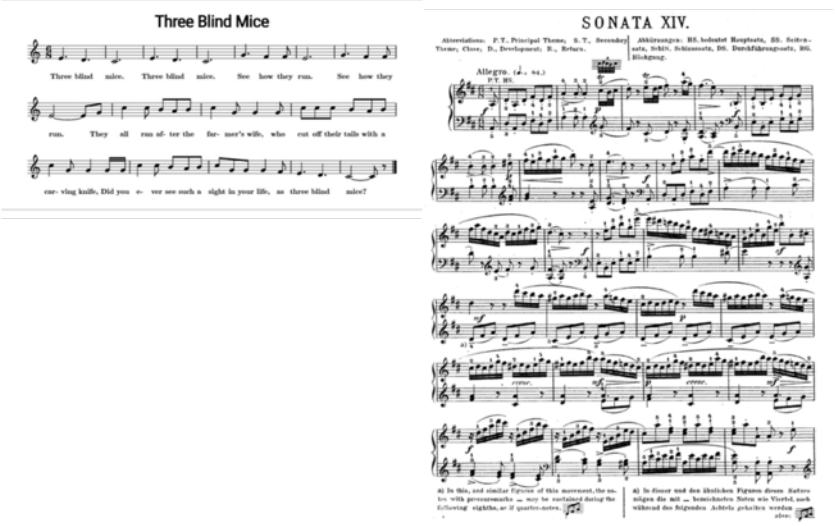
There are lots of metaphors to explain the difference between marks and grades, for example, in the images above the distance walked could be considered as marks, it is a common measure of how far someone has travelled, but in understanding how much of an achievement it was you need to consider where they were walking—this is taken into account in setting grades.

- Marks represent how much of the task a candidate has completed.
- A grade takes into account how difficult the task is to provide an indication of how impressed we should be by the candidate's mark.

Consider the following two examples. In the first example we would expect a 16-year-old to get nearly all of the task right to get a "good" grade, while in the second example we would expect far less to indicate they deserved a "good" grade.

Figure 35

We would expect less of good candidates when faced with a difficult task than an easy one



We would expect "good" candidates to have completed most of the task.

We would expect "good" candidates to have only completed part of the task.

This leads on to one of the challenges of setting good quality assessment. Candidates need to have the opportunity to show their full potential, which may not be possible if the tasks are too simple. Conversely, if the tasks are too challenging, we can have a situation where the least able 50% cannot even start the task so we cannot differentiate between them in our grades.

The other danger from having tasks that are too simple is that we may start measuring accuracy rather than understanding—so that candidates who make minor mistakes may be excluded from the top grades despite having a good comprehension of the topics. This is not what was intended for the assessment.

The tyranny of grades—lesser of two evils

Consider the knowledge, skills and experience that go to make up an excellent chef. The knowledge—of ingredients and flavour combinations that will make up a perfectly balanced dish. The skills—in selecting and preparing ingredients, of cooking on the hob or in the oven, of presentation. And the experience—of using technique to achieve perfection, of judgment to know when something is perfectly cooked, of presenting a combination of ingredients and dishes to achieve sublime satisfaction at the table.

Figure 20

Knowledge, skills and experience

Now reduce all of that complexity into a single grade out of 7 to decide who the best chef is. The result is almost meaningless. Does it help if we increase the scale to 100, or even 1,000, grades? The answer is likely to be that it doesn't help; there may be more scope to differentiate but the fundamental issue of trying to compare different skill sets does not go away.

This is exactly the issue faced within assessment; how to represent the complexity of a learner's knowledge, understanding and ability to synthesize into a single outcome. Even if we could precisely capture all the information about the candidate it would still not mean that we could give them a grade that perfectly reflects their talents.

The alternative suggestion, which is often proposed, is not to award a grade at all, perhaps to provide each candidate with a personalized description of what they did well in and where they were less strong. Such an approach is in keeping with the principles of good teaching and learning, but has the severe drawback (which some proponents of this approach regard as a strength) that comparisons between candidates are very difficult.

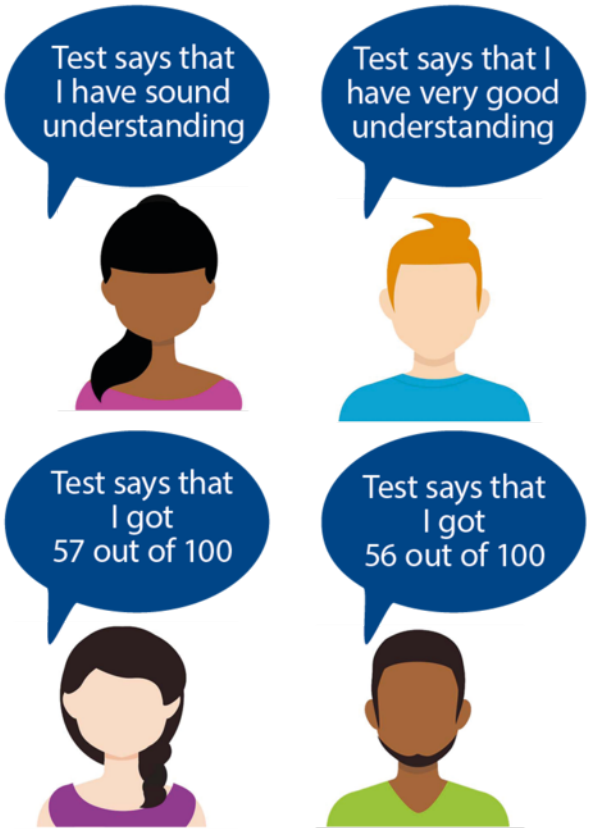
To attempt to resolve this problem we return to the purpose of the IB assessment. It is to support students in being able to progress to further study or work. This means that there will need to be some kind of selection process by the receiving institutions (often universities for DP/CP and schools for MYP) to determine which students can be given these opportunities.

> Clearly, if the other criteria are less reliable than the examinations, greater reliance on them will lead to less reliable selection decisions.
>
> (Cresswell 1986: 37–54)

This quote featured in the introduction to this guide and it bears repeating. If selection is going to take place, then the IB has a responsibility to support its students by making it as fair and meaningful a selection decision as possible. If we only provide descriptive accounts of students to the receiving institutions, then they will need to find some way of comparing which will almost certainly be less reliable and comparable than that offered by grading examination outcomes. Consider, for example, the validity of a short interview with a tutor and all the factors that could influence the outcome which should not be the basis of selection.

This is not to suggest that summative assessments are a perfect, or even a particularly good way of making such selection, but they are fairer than the alternatives, and most importantly the IB is constantly striving to make them as fair, meaningful and reliable a method as possible.

Figure 21: Should we differentiate between the two candidates in each pair?



In the example above, the difference between the first pair of candidates is taken from the generic grade 5 and grade 6 DP grade descriptors. If you had to make a decision between the two candidates it would probably be fair to do so, and it is likely that if they took the tests again they would get the same outcome. In the second pair, the difference is likely not to be meaningful, and the two candidates probably have the same ability.

What this example demonstrates is that not all differences are meaningful, and the use of grades provides an indication of where we feel we can differentiate between two candidates. It is certainly true that for a candidate on the boundary (either just above or just below), either grade is fair, but for most candidates a difference in grades represents a meaningful difference in performance.

This leads on to the debate of how many grades to have. If you have only two grades (pass or fail) then most candidates will get a fair grade but for those on the boundary the consequences are very serious. In contrast, if you have 20 grades far more candidates lie on a boundary, but for each candidate the consequence of being in the wrong category is less significant. This concept is explored in more detail in Cresswell (1986).

In the IB, we have generally selected seven grades as representing the number of meaningful categories our assessments can provide and also the right balance between the number of candidates on a boundary and the impact of being in the wrong grade.

What is a successful examination session?

Figure 36

What makes a successful session depends on your point of view

I knew the answers to all the questions.

Our candidates did better than nearby schools

We were able to pick the right students for our courses.

We were able to pick the right students for our courses.

What success looks like depends very much on your point of view, and each is equally important. In the IB, we focus on both high-level validity and practical delivery.

- The assessments allowed every candidate a fair opportunity to show their ability.
- The experience for the schools and candidates was as straightforward as possible.
- All our stakeholders retain confidence in the outcomes (grades) we have released.

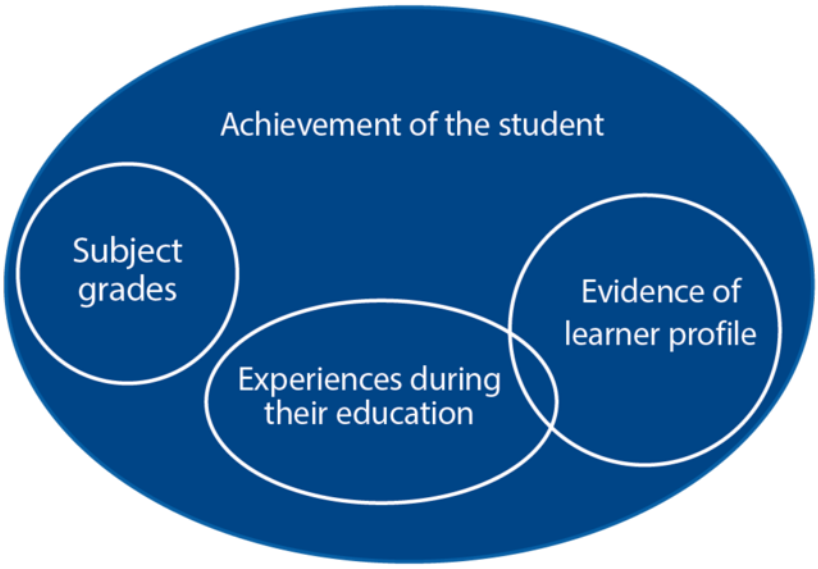
Achievement is more than just grades

As seen in the IB mission statement, the goal of an IB education is far more than a series of academic grades. This is reflected in the learner profile and articulated in *What is an IB education?*

Assessment outcomes can only focus on a narrow part of this mission statement. Given the compensation model used in our examinations, what proportion of marks in our mathematics assessment should be allocated to caring or risk-taking, even assuming that giving a "numerical value" to these important traits is reasonable? However, one of the principles of IB assessment is that it should have a positive backwash effect, which would then support these goals.

Figure 37

The achievement of a student is far more than can be evidenced



When reporting the outcomes of an IB programme it is important to consider more than just the assessment grades in order to reflect the full range of student achievement.

The other aspect of this is that the IB only records the final attainment of the candidate in the assessments without any indication of how challenging it was for them to achieve this result, or their full potential. We recognize the importance of both these elements, but believe it is not possible to measure them meaningfully within summative assessment. Such evaluation is properly the responsibility of the school who have a holistic view of the candidate.

While there are methods to calculate "value added" measures of candidate attainment or "predicted grades based on prior attainment" we would advise caution regarding their use as an indication of student success. Such measures are based on an average student, and each student is a unique combination of personal characteristics and traits which should be celebrated appropriately by those who have the opportunity to learn with them over a whole programme.

Assessment process: Roles and responsibilities

- Each of the roles in the assessment process has their own responsibilities and skill set.
- In some cases, it may be the same people fulfilling these different roles at different points of the assessment cycle.
- Some of these roles are fulfilled by the IB, and for others we draw upon experts from the IB community. In the latter case, the IB is responsible for the final sign off and maintaining the quality of these aspects of the assessments.

Figure 38

The accountabilities and responsibilities of the key players in the assessment cycle

Principal Examiner (PE)	**Chief Examiner (CE)**
• Responsible for one component • Final arbiter on what mark candidates' answers receive in that component • Ensures that all examiners understand the marking standard in that component • Guides CE in setting grade boundaries for the component	• Overview of all components in their subject (group) • Ensures consistency of standards between all components, including in paper authoring • Arbitrates on any academic issues relating to the assessment • Recommends final grade boundaries to the IB
Examiner	**IB**
• To mark candidate work in accordance with the standard set out by the PE	• Accountable for all aspects of assessment • Responsible for the assessment processes such as examiner recruitment, examiner quality and issue of results • To make decisions on issues of academic misconduct, or maladministration, or special arrangements and considerations • To accept or challenge the grade boundaries recommended to it by the CE

Principal Examiner and Chief Examiner

The Principal Examiner (PE) looks after one component (that is, internal assessment or paper 2) and the main tasks they are responsible for are:

- deciding what answers are awarded marks (setting the marking standard)
- explaining this standard to their team of examiners; the PE edits the markscheme and any additional guidance to the examiners, including the practice scripts
- guiding the Chief Examiner in setting grade boundaries for their component.

In summary, the PE is the person in charge of the academic issues around the marking of a particular paper. The PE is supported by a number of other experienced examiners (called senior examiners) who discuss issues and make recommendations to the PE.

It would be very unusual if the person who will be the PE for an examination was not part of the paper authoring team, but this is not a requirement of the role.

The PE is engaged by the IB for the examination session but is not a member of IB staff. Components will generally have the same PE for several sessions to help ensure consistency.

The Chief Examiner (CE) is responsible for maintaining the quality of several related components. In the DP and CP this means they are responsible for a whole subject, while in the MYP (which only has one component per discipline) the CE is responsible for a subject group, such as science or individuals and societies. They act as the IB's academic expert in this field and provide leadership to their PEs in settling any disputes.

The CE is responsible for:

- making sure that standards are appropriate within and between components
- ensuring consistency between components, both in marking and paper authoring
- arbitration in any disagreement between/with PEs
- leading the grade award process and recommending grade boundaries to the IB
- acting as ambassador on behalf of the IB.

The CEs are also invited to work with the IB in discussing and improving our assessments. Most CEs also act as PEs for one of the components they are responsible for so they have first-hand experience of candidates' answers in their subject area. Unlike PEs, the CEs also have a responsibility for ensuring the high quality of their subject papers during the authoring stage.

The CEs are not members of IB staff but are contracted to work with the IB for between two and seven years. The upper limit is in place to encourage progression to ensure that the IB assessments do not become stale.

In some small entry subjects, it is not appropriate for the IB to appoint a CE. In such cases, the IB will appoint an "Examiner Responsible" who will take the role of ensuring consistency between components and recommending grade boundaries to the IB. This person will also act as a PE for a component.

Usually, PEs and CEs would also be involved in the curriculum review cycle which is not dealt with in this document.

Other examiner roles

The examiner is responsible for marking candidate work to the standard set out by the PE. They need to prove they have understood and are applying this standard through the quality model (see the section on "Quality model").

Examiners apply to the IB to mark in a session and must be an expert in the subject they wish to mark (usually through being a teacher of it) as well as having experience of teaching students in the relevant age range. The IB checks these credentials with the referees the applicant has provided. Examiners are generally only offered one internal assessment/coursework component and one examination component to mark in. The marking period for these two components will not overlap. Examiners are paid per "live" candidate script. "Live" scripts do not include any qualification or seed scripts (see the section on "Quality model") as these are just demonstrating to the IB that they understand the marking standard.

Team leaders are particularly experienced examiners who the IB asks to support other examiners in understanding the correct marking standard. Team leaders support examiners through the qualification process and provide feedback if the seed scripts indicate they are drifting away from the necessary standards. The number of team leaders in a subject will depend on the number of examiners required. In a small entry subject the PE may provide this support to all examiners.

Senior examiners are experienced examiners who are asked to support the PEs in their various tasks. Senior examiners would usually also be team leaders.

The responsibility of IB staff

Each subject will be allocated to an assessment subject manager who will manage the whole assessment cycle and ensure that the relevant quality control procedures have been followed. Many subject managers are also experts in the subjects and so are able to support the work of the CE and PEs.

Separate teams within the IB manage requests for modified papers, special arrangements, special consideration, allegations of academic misconduct (including plagiarism), monitoring of examiner quality and calculation of moderation factors.

The Chief Assessment Officer, supported by the relevant Head of Programme Assessment and Head of Assessment Principles and Practices, is responsible for considering the recommended grade boundaries put forward by the CE for each subject, and either agreeing them or asking the CE to reconsider the recommendation.

Ultimately, the IB is accountable for all decisions made as part of the assessment cycle. The use of external experts supports us in producing fair, high quality assessments but the final accountability rests with us.

Roles in authoring examination papers

The key skills required for a PE are to be consistent in your marking standard and able to explain to other examiners what the required standard is.

The key skills required for writing a high-quality examination are very different and include:

- creativity to produce engaging and distinctive questions that nonetheless reflect the curriculum
- awareness and understanding of the different teaching and cultural approaches to produce a question that is free from bias
- language skills to produce a question that is clear and unambiguous, and which will remain so when translated into the other required languages
- clarity to produce a question which tests specifically what was intended (construct relevance) in a way that can be marked consistently (reliability).

Generally PEs and senior examiners possess both skill sets, and so the IB can benefit from the same people writing and marking assessments, supported by IB subject managers.

In addition to the creative roles, there are a number of other people involved in the production of the final assessments.

- Technical design editors—who are responsible for transforming the assessment questions into the finished format of the examination, either paper or on-screen.
- Translators—to convert the assessment into the other languages in which it will be taken.
- External reviewers—subject experts who will take the assessment as if they were a candidate, thus providing feedback on the overall length of the papers and any ambiguity or errors.

For more details on the way assessments are authored, see the section on "Examination paper preparation—development and quality".

Examiner hierarchy

- The PE is the final arbiter of what mark to award on a component. All other examiners have to follow this standard.
- The IB makes the assumption that those who have been involved with the PE in setting the standard have a better understanding of the PE's standard than those who have learned it from the standardization material and their team leaders. This means that marks from examiners closer to the PE in the examination hierarchy take precedence over those who are further away.
- "Senior" refers to closeness to the PE in the examiner hierarchy and not the length of time they have been an examiner or their teaching experience.

The principle behind the IB's marking is that the PE is the final arbiter of what is the correct mark for a candidate's work. While we recognize that there may be many different and equally reasonable views on what is the "correct" mark, it is not fair for a candidate to get a different mark depending on which examiner received their work and so we ask the PE to set the standard and ask every other examiner to follow it.

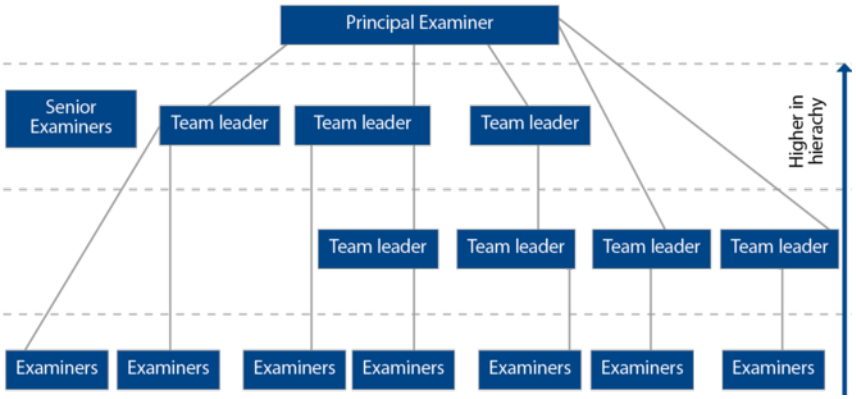
The PE is supported in setting the standard by a team of senior examiners. We make the assumption that, as these examiners have had the opportunity to discuss the standard with the PE, they will have the best understanding of it. Therefore, if they disagree with the mark given by another examiner (typically as a result of at risk re-marking or an EUR) then we will use the most senior examiner's mark as the final mark for that piece of work.

Most team leaders will be senior examiners, but where this is not the case, we hold a senior examiner's mark above that of a team leader. Similarly, we use a team leader's mark in preference to an examiner who has learned the PE's standard from the standardization material and their team leader's instructions.

This is what is meant by the hierarchy of examiners. For a particularly large subject we might have a structure like the following.

Figure 39

Hierarchy of examiner roles for a very large subject



The hierarchy is not inflexible. If, as a result of data from qualification or seed scripts, we discover that an examiner has a better grasp of the PE's standard than a more senior member of the hierarchy, then this will be taken into account—the role of team leader requires being able to explain the standard and not just to apply it.

In an ideal situation there is never a need to make decisions based on the hierarchy as everyone is marking to the same standard, but in reality it is an essential way of determining which marks the IB should use. If we come across a particularly difficult or controversial case, we would refer the script to the PE to determine the appropriate mark.

Integrity of the assessment

- IB assessments can only be fair if all candidates have an equal opportunity.
- The various forms of maladministration and academic misconduct create a disadvantage for those candidates who have followed the rules, and so the IB takes every effort to prevent such behaviour.
- The IB rules, set out in the *General regulations* and other documents for each programme, are designed to minimize the opportunity for academic misconduct or maladministration. Ultimately, it is only the school who can create the learning culture where academic misconduct is not acceptable and is reported.
- While certain forms of assessment are less susceptible to academic misconduct than others, the IB's principle remains that construct relevance (that is, testing what we really want to assess) should remain our primary consideration when designing assessments.
- On-screen assessment is likely to support us in our efforts to maintain the integrity of our assessments.
- If anyone has any suspicions about academic misconduct that is not being dealt with by the school they should contact complaints@ibo.org or contact IB Answers.

Academic honesty is a set of values and skills that promote personal integrity and good practice in teaching, learning and assessment. It should not be imposed as a series of strict rules, but should instead be a culture within a school and the wider community including legal guardians. While it is easier to explain to students what academic **dis**honesty is, with reference to collusion, plagiarism and cheating, such an approach will not create the kind of positive culture of integrity which will organically lead to fairer assessment outcomes.

In order for assessments to be valid, they need to provide an accurate reflection of a candidate's achievement relative to all the other candidates who have taken the assessment. For this reason, the IB takes great care in having consistent approaches to marking, grading and removing bias from its examinations. The rules and regulations it sets out are another aspect of creating this "level playing field".

The *General regulations* define academic misconduct as behaviour that results in, or may result in, the candidate or any other candidate gaining an unfair advantage. Such activity affects not only the candidates involved, but everyone who has taken the assessment as it reduces the validity of the qualification. The IB therefore takes academic misconduct very seriously, and details of its prevention and consequences can be found in the various programme *Assessment procedures* and publications on academic honesty.

Assessment design is an important tool in preventing academic misconduct as some types of assessments are easier to monitor than others. For example, in a written exam it is harder to obtain help from someone else than it is for an internal assessment piece of work. While we take such considerations into account in devising our approach to assessment, our principle is that we should not sacrifice construct relevance (that is, testing what we are really assessing) to prevent the opportunity for academic misconduct.

Manageability is another important aspect of validity however, and this must also be considered when preventing the opportunity for academic misconduct. Setting up examination halls, particularly with on-screen assessment, can be a real challenge for schools, and this must be taken into account. The IB endeavours to keep the experience of schools in mind when setting its rules and regulations and is keen to hear from coordinators and head teachers about good and challenging practices.

This communication also extends to understanding about new and emerging issues in academic honesty and anyone with concerns or thoughts is encouraged to contact complaints@ibo.org or IB Answers.

Dealing with conflicts of interest

Within the IB, access to examination papers is carefully controlled and the IB actively manages any connections to candidates taking our examinations which could constitute a conflict of interest. Staff responsibilities are reorganized if such a conflict could be seen to occur.

The IB principle is that no examiners can mark their own candidates' work. In the extremely rare occasions when this is unavoidable due to circumstances beyond the IB's control, for example there is no-one else qualified to mark at the required standard, a second independent examiner would review the marking to establish there is no indication that a different standard has been applied to the examiner's own candidates.

Managing maladministration

The IB *General regulations* and *Assessment procedures* documents set out rules and instructions which minimize the chance of maladministration occurring.

Examination papers must be kept in a locked safe in a locked room to prevent any unauthorized access. Papers are sent in tamperproof bags so that it is obvious if they have been opened. Examination paper breaches are some of the most challenging situations for the IB to mitigate because it is difficult to know how widely the papers have been shared, and so any action will usually affect a large number of candidates to ensure a consistent experience for all of them.

Examination papers packages must be opened in front of all candidates so they can see they have been kept securely. If this does not happen candidates should contact the IB.

Invigilators should remain vigilant throughout the examination to ensure no academic misconduct takes place.

The IB does carry out school inspections during assessment sessions to ensure that these practices are in place. However, these can only be spot checks on the processes and the IB places great responsibility for preventing maladministration in the hands of the heads of schools and programme coordinators, who are able to ensure high standards are maintained on a daily basis. Further, they should ensure that the culture in the school is one that encourages best practice and high levels of integrity from its teachers and students.

Challenges with international examinations and time zones

The IB faces some specific challenges around the international nature of its schools. Within most national systems, all candidates can sit the examination at the same time, but due to the cross time zone nature of the IB this would not be fair on some candidates. Imagine a situation where the examination started at 5am or finished after midnight.

This reality means that some candidates will have finished their examinations before other candidates have started them, and so we require integrity in our candidates and teachers not to take advantage of this. Where candidate numbers are large enough to make it viable we do have two separate time zone papers to reduce the time between candidates finishing the examination in one time zone and others starting in the other time zone, but even in these cases there can be a considerable variation in the start time of an examination.

There are a number of mitigations to address this. The first is simple self-interest—why would candidates choose to advantage other candidates over themselves? The second is careful monitoring of websites where papers could be shared. The rule that candidates cannot take question papers out of the examination hall is in place to limit the opportunity to share questions online.

The third mitigation is to ensure that questions are designed to test understanding rather than knowledge recall. Given the limited time scales available, the benefit of having a little more time to consider your answer, when speed of thought is not what the test is assessing, is of limited benefit when compared with

memorizing "recall" answers. The final mitigation is that by imposing a requirement for all candidates not to discuss the examination for 24 hours after the examination there is no possibility of "innocent discussion".

The IB continues to seek to innovate in this area to reduce the risks from time zone cheating.

Plagiarism in coursework

One of the most challenging areas to manage is ensuring that work undertaken in the classroom (and at home) is the candidate's own work.

The IB is aware of the range of websites offering to "support" candidates with their work, and our best defence against this kind of academic misconduct is the teacher who will have worked with the candidate and can identify where the work does not reflect the candidate's usual standard. For this reason, the candidate and teacher are asked to confirm that the work submitted is the candidate's own. This is not a simple added complication, as the IB is not obliged to accept an alternative version once the work is submitted.

Establishing and managing a culture of academic honesty is a requirement on all IB schools and repeated breaches will have consequences for authorization. Simply using Turnitin software is not sufficient: teachers should also work with candidates as they write their IA to check the authenticity of the work.

For the avoidance of doubt, no level of plagiarism is acceptable, and all citations from other authors must be properly referenced as set out in the IB regulations. The IB uses a range of software, including but not limited to Turnitin, to identify plagiarism.

Benefits of on-screen assessment

On-screen assessment offers a number of benefits in managing academic misconduct. It limits the opportunities for the papers to be accessed before the examination through the use of encryption and passwords. It also allows monitoring of when examination packages were opened, and by whom.

For the maladministration point of view, the IB can require schools to record and justify any modifications they make, such as extra time or pauses in the examination. The means the IB can devolve more responsibility to schools for reasonable modifications while ensuring fair practice.

The on-screen package also allows for more detailed understanding of when answers were given, for example, proving that a candidate completed a question before an alleged incident occurred. Also, the electronic nature of the answers allows for large-scale checking for similar answers which is not possible in handwritten scripts.

We recognize that on-screen assessments will be subject to new forms of academic misconduct, particularly hacking attempts, but tools are being developed with such attacks in mind.

Resources

Teachers, schools and students are best placed to challenge and stop cases of academic misconduct through creating a culture where it is not acceptable and by being vigilant in tackling it when it occurs. The following materials can support schools in creating a culture of academic integrity, and can be found on the IB website under "Digital toolkit".

- *Academic honesty in the IB educational context*
- *Effective citing and referencing*

If you would like further support or have any concerns about academic misconduct which is not being addressed at your school, please contact complaints@ibo.org or IB Answers.

Fairness for all—meeting candidates' needs

- In order for IB assessments to be valid they must not discriminate against candidates with particular needs. The IB will consider requests for modified papers and inclusive arrangements as set out in the programme's *Assessment procedures*.
- The best way to ensure fairness is through designing assessment to be accessible for everyone, thus removing the need for any modification. This is encapsulated in the concept of Universal Design of Assessment, part of the IB's commitment to Universal Design for Learning (UDL).
- Some candidates' needs are known about in advance and these are dealt with through inclusive access arrangements which may include modified papers.
- Other circumstances arise at short notice or cannot be managed through inclusive arrangements. In these cases, we treat them through our special consideration processes.
- Ultimately, the purpose of all these arrangements is to create fairness for all our candidates, and so in reaching any decision the IB must consider what is fair for the entire cohort and not just the one individual candidate. The aim is to have an even playing field for every candidate.

The IB believes that all candidates should be allowed to demonstrate their ability under assessment conditions that are as fair as possible. We recognize that standard assessment conditions may put candidates with learning support requirements at a disadvantage by preventing them from demonstrating their level of attainment. Similarly, we acknowledge that sometimes events or circumstances beyond the control of the candidates will affect their performance and should be taken into account.

The best way to ensure fairness with an assessment is for everyone to take the same assessment in the same way. Many of the modifications made to support specific requirements would help all candidates in understanding and engaging with the questions. The ideal situation is for all assessments to be developed with an understanding of the range of requirements that candidates may have rather than to treat some candidates differently. This is the concept of Universal Design of Assessment. The IB recognizes that this total inclusivity approach is sometimes not achievable and so we also have a process for requesting specific inclusive arrangements.

Inclusive access arrangements are designed to meet candidates' individual needs, such as:

- learning disabilities
- language difficulties
- specific learning difficulties
- communication and speech difficulties
- autism spectrum disorders
- social, emotional and behaviour challenges
- multiple disabilities and/or physical, sensory, medical or mental health issues.

Any reasonable adjustments for a particular candidate pertaining to his or her unique needs will be considered. For further details, please refer to *Assessment procedures* and the IB publication *Candidates with assessment access requirements*.

Adverse circumstances are those that are beyond the control of the candidate and which might have a negative impact on their performance. Such cases are considered by the Final Awards Committee and if accepted candidates close to a grade boundary will receive the higher grade.

The accepted IB principle of fairness to all candidates means that, when considering any inclusive arrangement or adverse circumstance, we should not create a situation that is unfair for other candidates taking the assessments. The goal is for a level playing field for all candidates.

Principles for inclusive access arrangements

The principles for inclusive access arrangements are set out in the IB document *Candidates with assessment access requirements*. The text below is taken from the DP version, but similar principles apply to other programmes.

1. The IB must ensure that a grade awarded to a candidate in any subject is not a misleading description of that candidate's level of attainment, so the same standards of assessment are applied to all candidates, regardless of whether or not they have learning support requirements.
2. Inclusive access arrangements, including reasonable adjustments, are pre-examination measures for a candidate to access the assessment. They cannot be requested retrospectively either for oral or written examinations.
3. The arrangements requested for a candidate must not give that candidate an advantage in any assessment component.
4. The inclusive access arrangements described in this document are intended for candidates with the aptitude to meet all assessment requirements leading to the award of the diploma or course results.
5. When inclusive access arrangements are necessary for a candidate during the course of his or her study of the Diploma Programme or practice examinations, the school may provide the arrangements. If the arrangements are required for assessment, this document lists the arrangements that do not require prior authorization from the IB. For all other arrangements, prior authorization from the IB Global Centre, Cardiff is mandatory. Similarly, if a Diploma Programme candidate has difficulties meeting the requirements for creativity, activity, service (CAS), IB Answers must be consulted.
6. Schools are advised to plan inclusive access arrangements for their candidates based on the IB criteria as stated in this policy and teachers' observations of the candidate in the classroom during class work and tests.
7. The inclusive access arrangements requested for a candidate must be his or her usual way of working during his or her course of study. Only in very exceptional and unusual cases, will the IB authorize a request for inclusive access arrangements that are not the usual way of working and that have been put in place to support the candidate only in the last six months of study or thereafter, just prior to the examinations.
8. The IB aims to authorize inclusive access arrangements that are compatible with those normally available to the candidate concerned. However, authorization will only be given for arrangements that are consistent with the policy and practice of the IB. It should not be assumed that the IB will necessarily agree to the arrangements requested by a school. Coordinators are required to provide information on the candidate's usual method of working in the classroom.
9. The IB is committed to an educational philosophy based on international-mindedness. Therefore, the inclusive access arrangements policy of the IB may not reflect the standard practice of any one country. To achieve equity among candidates with assessment access requirements, the policy represents the result of a consideration of accepted practice in different countries.
10. The IB will ensure that, wherever possible, arrangements for candidates with a similar type of access requirement are the same. Due to the cultural differences that occur in the recognition of learning support requirements and the nature of access arrangements granted in schools, there may be some compromise that may be necessary to help ensure comparability between candidates in different countries.
11. Each request for inclusive access arrangements will be judged on its own merit. Previous authorization of arrangements, either by the IB or another awarding body, will not influence the decision on whether to authorize the arrangements that have been requested by the coordinator.
12. The IB treats all information about a candidate as confidential. If required, information will only be shared with appropriate IB personnel and members of the final award committee, who will be instructed to treat such information as confidential.

13. If a school does not meet the conditions specified by the IB when administering inclusive access arrangements or makes arrangements without authorization, the candidate may not be awarded a grade in the subject and level concerned.
14. If it can be demonstrated that a candidate's lack of proficiency in his or her response language(s) arises from an identified learning support requirement, inclusive access arrangements may be authorized. (For subjects in groups 3 to 6, all candidates are allowed to use a bilingual/translation dictionary in the written examinations.)
15. A school must not inform an examiner of a candidate's challenges (such as autism, writing difficulties and so on) or adverse circumstance.
16. In the case of internally assessed work, teachers must not make any adjustments when marking a candidate's work.
17. The list of inclusive access arrangements available is revised regularly. The IB will consider alternative arrangements proposed by a coordinator, provided those arrangements could be made available to all candidates with similar requirements.
18. According to the document *General regulations: Diploma Programme*, a Diploma Programme candidate may participate in three examination sessions to be awarded the diploma. At the discretion of the IB, a candidate with learning support requirements may be allowed additional sessions.
19. If the nature of a candidate's challenge and/or the authorized inclusive access arrangement might disturb other candidates during an examination, the candidate must take the examination in a separate room and be supervised according to the regulations governing the conduct of Diploma Programme examinations.
20. Written examinations must be invigilated according to the regulations governing the conduct of Diploma Programme examinations. The person invigilating the candidate's examination must not be a relative of the candidate, or any other person with whom there may be an apparent or perceived conflict of interest.
21. Any issues that arise from the nature of the inclusive access arrangements, or any unforeseen difficulties encountered by the candidate during the examinations, should be reported to IB Answers as soon as possible.

Exemptions from assessment

Exemptions are not normally granted for any assessment component. However, if an assessment component or part demands a physiological function that a candidate is not able to perform, an exemption may be authorized. Before submitting a request for an exemption from a component, careful consideration should be given to whether all reasonable adjustments have been considered. Authorization for an exemption will only be given when there are substantial grounds for an exemption. A candidate's physical inability to perform the functions required by the component must be clearly and fully documented.

For full details on the principles and processes around exemptions from assessment please refer to the *Candidates with assessment access requirements* document for the appropriate programme.

Opportunities for inclusive arrangements with on-screen assessment

The use of on-screen assessment allows the candidate to take far more control over how they wish the assessment to be presented. Computers are able to provide a large variety of fonts, text sizes and colours to meet an individual's needs and this type of adjustment can be routinely available to every candidate.

In many current cases, the inclusion arrangement requested is to allow the use of a computer and this need is clearly met by eAssessment so long as the inclusion software required is compatible with the on-screen tool. The IB is very aware of this requirement and is working to ensure that any on-screen examinations meet the industry standards for such inclusion software.

Adverse circumstances

Adverse or unforeseen circumstances are those that are beyond the control of the candidate and which might have a negative impact on his or her performance. This includes temporary illness or injury, severe stress, exceptionally difficult family circumstances, bereavement, or events that may threaten the health or safety of a candidate. Adverse circumstances may also include an event that affects the whole school community, such as civil unrest or a natural disaster.

Adverse circumstances do not include shortcomings on the part of the school. It is a school's responsibility to ensure that all candidates comply with programme and assessment requirements, including issues with teaching staff.

Full details of what is included and excluded within the category of adverse circumstances can be found in the appropriate programme's *Assessment procedures* and *General regulations*.

In such cases the evidence, supplied by the school, will be considered by the final award committee to determine if the candidate affected should be eligible for special consideration. If a candidate's circumstances are deemed "adverse" and therefore qualify for consideration, an adjustment may be made to the candidate's total mark in the affected subjects or programme core requirements. If the candidate is within one or two scaled marks of the next higher grade boundary, the candidate's grade in the affected subjects will be raised.

Universal design of assessment

The preceding text has discussed how the IB manages situations where candidates need modifications or specific assistance to be able to fairly take our assessments. The best solution, however, is to have assessments which do not have these barriers to participation in the first place. The concept of Universal Design of Assessment is to consider access, inclusion, equality, cultural sensitivities, stereotypes and bias from the starting design of the assessment. This includes the creation of examination tasks and questions, but also goes a step back into the design of the overall assessment model which sets the framework of how comparable assessments are created for every session. By creating more inclusive, and indeed less construct irrelevant, assessments at the start we can minimize the challenges faced in meeting the needs of individual candidates.

Universal Design of Assessment is an aspect of the overall Universal Design for Learning (UDL). UDL focuses on creating accessible learning environments for all learners, including candidates with disabilities, candidates from culturally and linguistically diverse backgrounds and candidates who are gifted and talented. The principles of UDL are relevant across the education including in curriculum design, school management and teaching. For more details on UDL in the IB, refer to Rao K, Currie-Rubin, R and Logli C. 2016. *UDL and Inclusive Practices in IB Schools Worldwide*.

The assessment cycle

- The process of creating assessments should be thought of as a continuous cycle whereby each stage is informed by the previous stage and leads into the next stage.
- It takes on average 18 months to create an examination paper (and its mark scheme), therefore the IB creates exam papers for different sessions in parallel with each other.

The life of an assessment can be thought of as a cycle, from its creation, through sessions taken by candidates, marked by examiners and the results released. The critical part of the process is that we learn from one session to improve the quality of the assessment for the next.

The diagram below is only one way of describing the process, many of the steps could be separated out or combined in different ways, but it is a good way of explaining the life cycle of an assessment. The table below summarizes each step and provides a link to the relevant section.

Figure 40

The assessment cycle

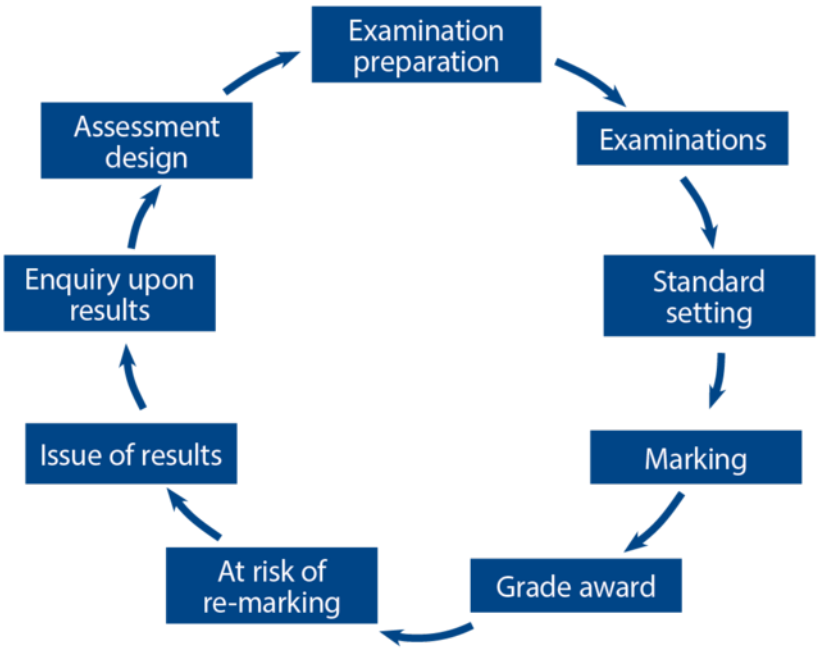


Figure 41

The assessment cycle

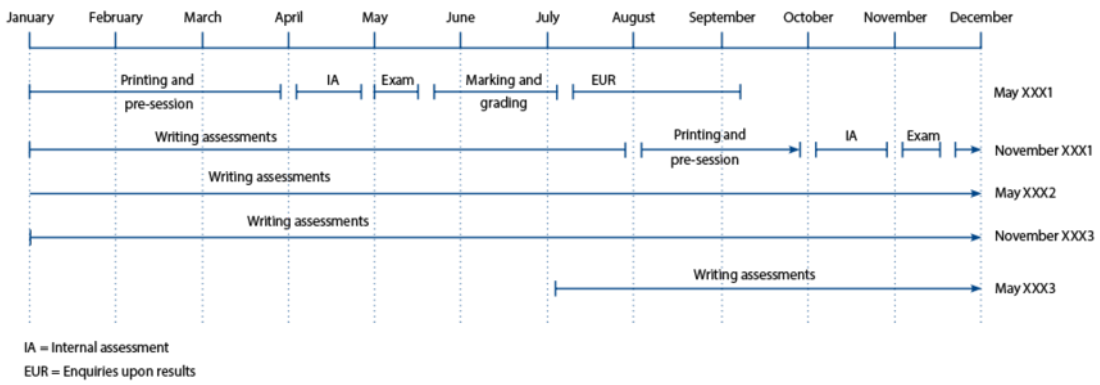
Stage of assessment cycle	Description
Exam preparation	The process of creating each individual examination. It covers everything from deciding on the topics the questions will cover, through writing and

Stage of assessment cycle	Description
	editing the specific questions, translating into other languages, arranging and preparing them in the correct format, and finally doing the necessary quality checks.
Examinations	This part of the process is where the candidates take the assessments in schools.
Standard setting	This is the process where the senior examiners explain how to mark candidates' work to their examiner teams, and identify the "definitively marked" scripts which will be used to check examiner quality.
Marking	This involves examiners looking at individual candidates' work and deciding how many marks to give it. They must follow the instructions set by the principal examiner and are regularly tested to make sure they are doing so correctly.
Grade award	Where our senior examiners decide how marks (which depend on the exam) should be converted to grades (which always mean the same).
At risk re-marking	The final quality check on the marking. It focuses on any areas we have evidence that an issue may exist.
Issue of results	The release of results to schools and candidates. This also covers mainly administrative processes like determining whether the course results mean the candidate has passed the programme, and sending transcripts to universities.
Enquiry upon results	The opportunity for schools to highlight where they think there has been an error in the examination process and ask the IB to look again at the candidate's work.
Assessment design	The most important aspect of closing the assessment cycle. To learn from the experiences of candidates undertaking examinations and to improve what we are intending to assess and the approach we take, including the number and type of assessment tasks.

To give you an indication of how long a paper takes to develop, the diagram below gives an indication of the time spent on each part of the assessment cycle. The cycle represents three complete years and highlights that on average it takes 18 months to prepare an examination paper. It also highlights that the IB will be developing both the May, November and following year's paper at the same time.

Figure 42

Assessment cycle



Impact of eAssessment on the assessment cycle

The introduction of eAssessment does not change the principle of the assessment cycle. It will allow us to be quicker in certain parts of the cycle (for example, it removes the need for sending examination papers and scripts to scanning centres). However, each part of the assessment cycle will still need to be completed.

Examination paper preparation—development and quality

- The outcome of this stage is that the complete assessments are ready to be taken by candidates.
- There are several stages of finalization that an assessment must pass through during preparation:
  - content (paper authoring)
  - layout
  - usability
  - translation
  - delivering the assessment to schools.
- Any modifications required by candidates (access arrangements) are also considered during this stage.

Rules underpinning the writing of the examination papers

The qualities that make a good assessment are described in the section on "assessment principles"and these qualities underpin the IB's work in writing examination papers. It is also important that the purpose to which the assessments will be put is kept in mind. The following six overarching rules can be considered as the key elements we expect our authors to keep in mind.

- Authenticity and construct relevance are the most important aspects of validity.
- Assessment should provide opportunities for candidates to demonstrate what they can do, not identify what they cannot do.
- The assessments need to be accessible to the widest possible range of candidate expertise, but also allow for reasonable differentiation.
- The best assessments are accessible to all; and we aim for inclusive arrangements to be unnecessary (Universal design of assessment).
- The assessment must test only the curriculum as set out in the subject guide.
- Consider how the assessment will be marked while it is being developed to ensure that what you are trying to test is what is awarded credit.

Explicitly this means that the markscheme should be developed along with the examination papers and that guidance on the anticipated tolerances for examiners are included.

Command terms

Command terms are key terms and phrases used in the syllabus content and in examination questions to indicate what is required in response to a particular command. This also suggests the type and depth of response that is expected.

While command terms have a common meaning across all subjects, it is important as a part of paper editing to ensure that they are being used as described in the relevant IB subject guides as students are expected to develop an understanding of these key terms in the context of their particular subjects.

Command terms present a particular challenge when they are translated into other languages as the subtleties of meaning can often be lost. To prevent this happing the IB publishes its command terms in each response language and explains their meaning. While this can sometimes lead to particular words being used in a linguistically unusual way, it is less problematic than having ambiguity of meaning in the

assessments. A key responsibility when checking translated papers is to ensure that the correct command term has been used.

Command terms broadly follow established taxonomies such as Bloom's taxonomy of educational objectives (Bloom et al 1956).

Question banks

The IB currently does not use any kind of question banks. Each paper is drafted as a holistic assessment by its authors.

In some other assessment systems, questions are extensively trialled to determine their cognitive demand, effectiveness as a discriminator, and so on and the question bank holds this information as well as more descriptive details such as how the question relates to the curriculum and number of marks, and so on.

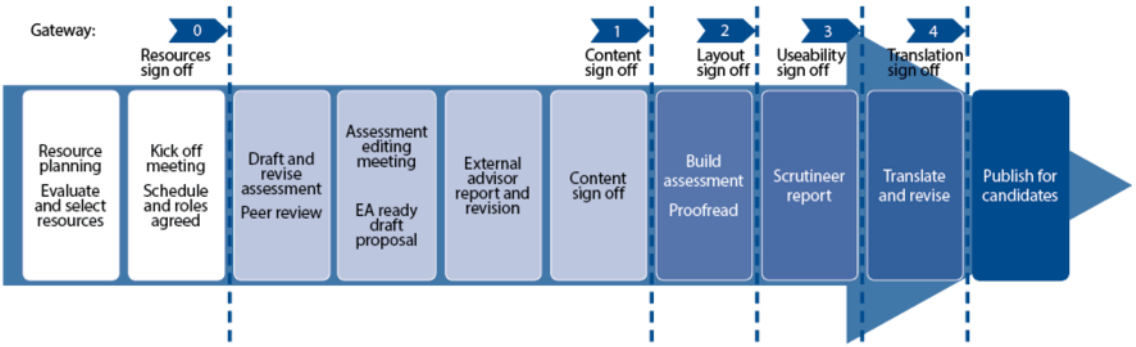
The IB does not do such trialling of its assessments because of concerns about the questions being leaked into the public domain and because the nature of most IB questions means that they cannot be evaluated without a very extensive cohort of students.

It is likely that in the future the IB will move towards its own model of question banks, perhaps drawing upon the extended IBEN community for contributions. This would be in the context of our principle of emphasizing authentic and construct relevant assessments.

Overview of examination paper preparation

Figure 43

Examination preparation process



The diagram above describes the different stages of the examination preparation process. Creating an assessment is a lengthy process which takes about 18 months and involves a wide range of external experts and IB staff including paper authors, assessment experts, editing staff, external scrutineers and translation staff.

It is very important that the markscheme is developed alongside the examination. This ensures that the questions and marking are fully aligned in what they are trying to test.

Process up to resources sign-off

The first stage of the process is to select who will be asked to write the papers. While it is usually the Principal Examiners who are asked, we also take into account any factors that might suggest a conflict of interest such as school connections or responsibility for running training workshops. Typically, there will be one author for each paper, and these authors will then peer review each other's papers in the subject. For very small entry language subjects it may be necessary to have a single author for all papers.

This then leads to the "kick-off meeting" where everyone who is involved in the process meets to agree responsibility for each task and deadlines. These roles are:

- Chief Examiner (may well also be a paper author) (External)

- Author for each paper in the subject—so typically 3–5 in Diploma Programme subjects (External)
- External advisor (External)
- Scrutineer (External)
- Production editor (IB staff)
- Subject manager (IB staff)

Where the IB produces papers for two different time zones, they go through a completely independent paper production process.

Process to content sign-off

After the kick-off meeting, each author is required to produce a first draft of their paper and markscheme, which are then shared for review with the authors from the other components in the subject. This then leads to a formal assessment editing meeting, where each paper is reviewed in terms of:

- level of cognitive demand of the questions, and overall range of difficulty of the paper
- checking for any possible bias
- does the paper follow the best practice of accessibility for all?
- how well the questions cover the curriculum
- confirming that the papers match the published assessment model
- how does the paper relate to past papers and sample material?
- checking that the questions are not the same as example questions in any published resources.

This meeting will produce another set of completed drafts which will then be reviewed by the external advisor.

The external advisor is another subject expert who has not yet been involved in the writing of these papers and markschemes. They provide a "fresh pair of eyes" and are asked to comment on the level of difficulty of all the papers taken as a whole, how they compare with previous papers and how well they match the curriculum. The external advisor offers suggested changes which the author and subject manager then review.

The final content sign-off of the examination paper is given by the author, with the IB subject manager confirming that the paper has been reviewed properly following our own quality standards.

The markscheme is not signed off at this point, it is reviewed during standardization in the light of candidate responses.

Process to layout sign-off

Once the content has been signed off, the text of the questions is fixed. The IB production editor will then put the questions in the format in which the examinations will be sat.

For traditional paper examinations, this means adding the front cover with the instructions or rubric and applying the appropriate style rules as well as adding the various bar codes and blank pages to form the question booklet. This process also includes redrawing any diagrams to the necessary quality for printing.

With on-screen assessment, the production editor will need to create the examinations in the development environment, which is known as IBeADS (IB eAssessment Development System). This includes the instructions/rubric and appropriate styles.

In both cases the completed assessments are proofread against the signed off content.

Process to usability sign-off

The scrutineer is another subject expert who has not been involved in the production of the assessment before this point. They take the assessment as if they were a candidate, including considering the time taken to answer the assessment. The purpose of this check is to spot any errors in creating the final examination and identify whether there is any ambiguity in the instructions or questions.

The scrutineer also reviews the markscheme once they have taken the assessment.

In some small entry language subjects, this scrutineer role is done by a second native speaker of the language rather than a literature expert.

Once the formal feedback from the scrutineer has been considered, the assessment is finalized and ready to be printed (for paper-based assessments) or bundled into the examination package—"wrap and deploy"—(for on-screen assessments) and then sent to schools.

Quality control

It is essential that examination papers are produced without errors or ambiguities as such issues can have a significant impact of the candidates taking the examinations. The IB takes such issues very seriously.

In the processes that have been described, it can be seen that, at each stage, there is a quality checking process as well as the formal sign-off process, namely, the external advisor, the proofreader and the scrutineer. Experience has indicated that it is better to have a single check rather than multiple checks for the same error—the knowledge that a paper will be checked again by someone else appears historically to have undermined the strength of quality checks.

The principle the IB adheres to in its paper production is "get it right the first time and check it once".

Translations

Candidates should be able to take examinations in the language in which they have been taught, and the IB offers candidates the chance to take (non-language) examinations in a range of response languages. Currently, most subjects are available in English, French and Spanish. Based on local agreements, other specific subjects are offered in German, Japanese or other languages.

Translating assessments is not a simple task. It is critical that the process of translation does not change the meaning of the questions so candidates are neither advantaged nor disadvantaged by taking the assessment in a particular language. This is challenging because of the need to rephrase sentences when converting into a different language. Particular words also have different meanings in different languages, which can create or remove meaning.

The need to translate examinations is kept in mind from the earliest drafts of papers, and our authoring teams are well aware of the issues that can emerge. The formal translation process occurs at the end of the process after the usability sign-off. The translators employed by the IB are subject specialists to ensure the technical language in the assessment remains accurate. The translation is then compared with the original examination (not just the text) by a bilingual assessor as a final quality check.

Managing requests for modified papers

The best situation is that assessments are designed so they are suitable for all candidates. However, there are certain requests and requirements which cannot be managed through improving the design of the paper. Some modifications, such as a different font, different coloured background or enlarged format, are relatively easy to deliver and are managed during the publication for candidates stage at the printers.

Other modifications, such as Braille papers, take more preparation. They may also require the formulation of related, but different, tasks to gather evidence of the candidates' expertise in an area. A common example of this is when the question asks the candidate to comment on an image, which is not appropriate for a blind candidate. We then need to consider whether a description of the picture is a reasonable alternative, or whether this increases the cognitive load on the candidate (too much information) or provides too much guidance (the description focuses on the aspects that will gain credit). Alternatively, should the question be rewritten in a different way?

These decisions are made by the paper author, guided by the IB's access and inclusion manager. Where appropriate, we also consult with outside experts. While decisions are made on a case-by-case basis, the outcomes and reasons for the decision are recorded in an auditable form to allow quality control for consistency. Where particular special modifications are made, they need to be done in such a way that minimizes the differences in what is being assessed.

Because of the comprehensive discussion involved in these decisions, ideally they should take place before the content sign-off stage. This ensures that the original purposes of the questions are fresh in the mind of the author and also allows for the possibility of improving the original question to remove the need for any modifications. This requires that schools inform the IB of any inclusive access requests as early as possible.

Moving to an on-screen assessment

Creating on-screen assessments does generate some considerable differences in the way in which assessments are prepared, but the overall process outlined in the diagram remains the same. Indeed, this is the process that is already followed with the MYP on-screen assessments.

The most significant change is in how papers are developed. As on-screen assessments are introduced, authors will need to understand their potential and opportunities to make the best use of the new tools that will be available. This step-change in what can be assessed and how is one of the most exciting elements of on-screen assessment—but also one of the most challenging.

Production editors working on on-screen assessments will need a different and broader range of skills. Rather than copying text into the required examination template and checking the style requirements, they may need to develop media clips and ensure that the overall assessment works in the way the authors intended. The IBeADS system that is currently used for MYP provides a framework for eAssessment that is similar to the traditional template but has its own special requirements for editors to master.

The final significant difference will be in the way the assessments are delivered to schools. For more details see the *Guide to the MYP exam session*.

Examinations

- The purpose of examinations is to allow candidates to show their full capabilities in a controlled environment that offers a consistent experience for all candidates and minimizes the opportunities for academic misconduct.
- In setting the examination session the IB need to balance the needs of all candidates with manageability for the schools and completing the marking as quickly as possible to meet candidate expectations.
- The IB publishes clear rules for examination room behaviour to minimize the opportunity for academic misconduct. These are updated in response to new technology and changing environments.
- The examination timetable is a compromise between many conflicting priorities and globally represents the least worst option.
- Details for dealing with unexpected events, inclusive access arrangements, adverse circumstances and rescheduling can be found in the appropriate programme's *Assessment procedures*.

The period when candidates are taking their examinations is very demanding and nerve-racking for them. Our overarching principle for this period is to minimize the stress we place on candidates by:

- limiting the length of the exam sessions
- where possible, avoiding clashes with other exams the candidates may be taking (for example, national tests).

This must be balanced against the need for having sufficient assessment to be able to make valid conclusions on the candidate's performance and preventing the opportunity for academic misconduct.

Preparing and managing the examination hall

Schools must conduct examinations according to a strict set of regulations laid out in ***The conduct of IB Diploma Programme examinations*** (date as per session) available on the programme resource centre>Implementation>Assessment processes and procedures. These regulations cover everything from dealing with the receipt of examination material, through conducting the examinations, to sending the scripts off to a scanning centre.

We attach great importance to maintaining the security of the examination papers and the proper conduct of each examination because, if there is a perception with schools and stakeholders that academic misconduct has occurred, then the value of the candidates' results will be greatly diminished. In order to maintain global confidence, the IB investigates all reports of maladministration carefully, and random inspection visits are paid to schools by IB staff or consultants during the examination session to check on the security of the examinations and to ensure they are conducted according to IB regulations. In reality, these visits can only be spot checks on the processes; and the IB places great responsibility for preventing maladministration in the hands of the heads of schools and programme coordinators, who are able to ensure high standards are maintained on a daily basis. Further, we expect that the culture in the school is one that encourages best practice and high levels of integrity from its teachers and students.

One of the worst possible outcomes for candidates is to take the examinations and then for the school or postal services to lose their work. Where this happens, the IB can try to mitigate the impact on the candidate by estimating a mark, but this is only possible if there is other evidence (papers) to draw upon. To help ensure that this is the case, we require that the different answer papers (scripts) are sent to the scanning centre on different days, reducing the possibility of all the scripts being lost in transit.

The introduction of on-screen assessments further mitigates this risk as there will be copies of candidate work available from each part of the upload process.

Developing the examination timetable

As a result of wanting to minimize the length of a session we do allocate different subjects to the same time in the exam schedule. While we make every effort to consider the subject combinations candidates can take, we recognize that a small number of clashes are inevitable and procedures for dealing with these are described in the programme's *Assessment procedures*.

Examinations are scheduled to avoid more than six and a half hours of papers in a single day where possible. The normal pattern for the examinations with multiple components relating to a particular course is to schedule the two or three papers consecutively, starting one afternoon and finishing the next morning. Although not always possible, this arrangement is preferable to presenting all the examinations for a given course on the same day.

For most candidates, this spreads the examinations more evenly over the three-week schedule. It allows candidates the opportunity to recover overnight if they feel they have not done themselves justice in a particular examination.

Principles of designing the examination timetable

It is not always possible to meet all of these principles, and in such cases a compromise needs to be achieved. The IB publishes the examination schedule at least one year before examinations will be taken.

The following points are a high-level summary of the principles that underlie the creation of the timetable.

- It is not currently possible to take into account public, national or school holidays, or religious festivals because of the number of countries in which the IB programme is offered.
- Although it would be desirable not to hold examinations on either a Thursday or Friday out of respect for schools in the Middle East whose weekend falls on these days, in practical terms it is not currently possible to do so.
- Where there are subjects with particular regional or cultural links we will endeavour to take these into account, for example Arabic literature/language examinations will not be scheduled on a Friday.
- The IB uses registration data regarding subject combinations to ensure that the minimum number of candidates globally are impacted by subject timetable clashes.
- Candidates should not be expected to be examined in two different foreign (not response) languages on the same day.
- In courses with multiple papers, to minimize the risk of an unexpected event disadvantaging a candidate in all their components, they will be scheduled over at least two days.
- Candidates should be allowed to focus their revision for each subject on a tight window, so where papers occur on more than one day they will be scheduled on consecutive days where possible.
- Candidates should have their examinations spread over the whole of the examination period, rather than over a short period. This means that when possible, subjects with the highest candidature (for example, history, English, mathematics) will not be scheduled for consecutive days.
- For the same reason, the IB will attempt to schedule language and science examinations in each of the three weeks of the examination schedule.
- The IB's internal examination processing requirements mean that certain subjects will normally be early in the schedule; in particular, large entry subjects.

Dealing with unanticipated events

The *Assessment procedures* sets out how to deal with most unanticipated events. If in doubt contact IB Answers who will be able to provide advice.

Inclusive access arrangements and adverse circumstances

These topics are dealt with in detail in the "Fairness for all" section.

The IB believes that all candidates should be allowed to demonstrate their ability under assessment conditions that are as fair as possible. Standard assessment conditions may put candidates with learning support requirements at a disadvantage by preventing them from demonstrating their level of attainment. Inclusive access arrangements may be authorized in these circumstances, which may be special arrangements with taking the examination or modifications made to papers.

Events that occur during that examination session, which are beyond the control of the candidate, and are likely to have a negative impact on their performance, are known as adverse circumstances. These circumstances include temporary illness or injury, severe stress, exceptionally difficult family circumstances, bereavement, or events that may threaten the health or safety of a candidate. Adverse circumstances may also include an event that affects the whole school community, such as civil unrest or a natural disaster. They do not include shortcomings on the part of the school. In such situations, the school should contact the IB with details of the circumstances so the Final Awards Committee can decide whether they should be taken into account when determining a candidate's grade.

For details about these arrangements please refer to the relevant programme's *Assessment procedures*.

Rescheduling of examinations

Rescheduling of an examination poses a major risk to the integrity of the examination process. It means candidates will sit the examination a considerable time before or after other candidates take it and could very easily lead to academic misconduct. Examinations will never be allowed to be rescheduled a day before they were due to be taken, as the risk of a breach that would affect the majority of candidates taking the examination is too great.

There are three circumstances only in which the IB will authorize a candidate to take one or more examinations at a time and/or date different to the IB examination schedule:

- Conflicts between IB examinations scheduled for the same time and date
- Conflicts between the scheduling of IB examinations and the examinations of other awarding organizations, including university entrance examinations
- Emergency situations

Rescheduling requires that the school coordinator can guarantee the security of the examination. If rescheduling is authorized for an earlier or later time during the same day, the coordinator must ensure that the candidate(s) concerned will be supervised during the entire period between the scheduled and rescheduled time. This is to ensure there is no communication with any other candidate who has already taken the same examination.

The process for requesting a rescheduling of an examination is dealt with in detail in the programmes' *Assessment procedures*.

Standard setting—Preparing examiners for marking

The key purposes of standard setting are:

- for the PE to set the standard for the assessment (through consultation with fellow examiners) based on candidates' work
- to test and refine the markscheme to ensure it will allow fair reward of candidate effort
- to produce definitive mark scripts (practice, qualification, seed)
- to disseminate and share understanding with all examiners
- to confirm understanding of standard through examiners' "qualification" process.

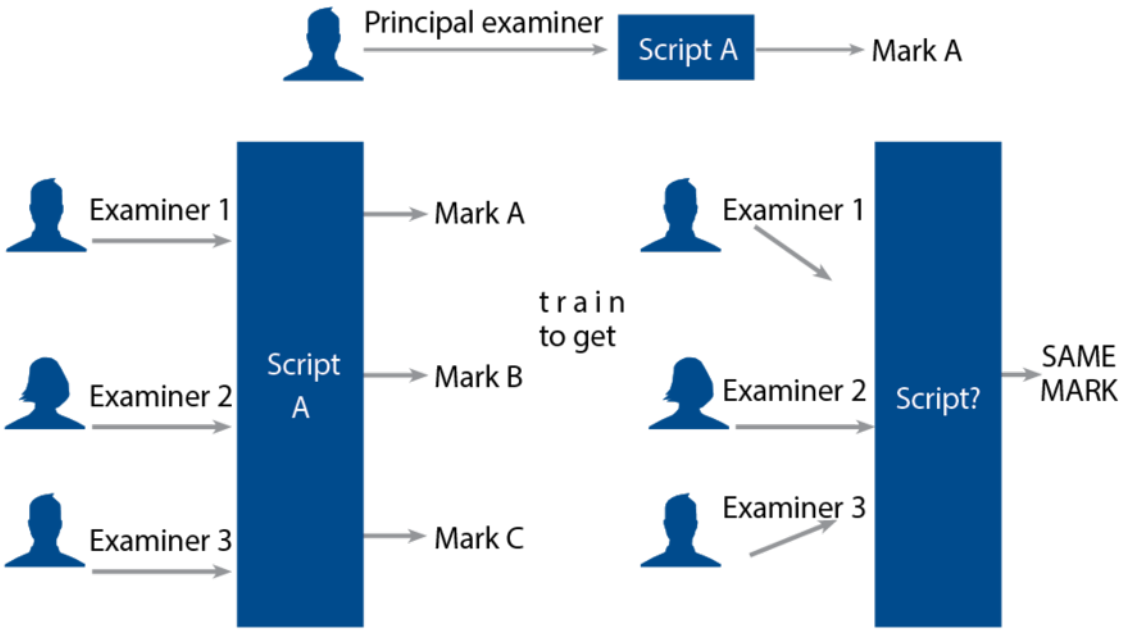
Standard setting is the period of time that covers several parts of the assessment process. The purpose of this stage is to prepare examiners so they can mark effectively and reliably.

It is important that every examiner will give the same mark to the same quality of work, otherwise the random chance of which examiner marks a candidate's work influences the final grade. The PE for a component sets this standard and every other examiner matches it.

This process of setting the same standard between every examiner is also known as "standardization".

Figure 44

All examiners must be marking to the Principal Examiner's standard



The first step in the process of standardization is to recruit examiners with the appropriate background and skills and then to provide them with general training so they understand the process and principles. More details of our examiner recruitment and training can be found on the IB website.

Standardization meeting

The standard setting discussion within the senior examiner team, led by the PE, is known as the standardization meeting. While it can be face to face, in most cases it is done virtually using a combination of IB discussion boards and video conferencing to remove the need for our senior examiners, who are spread around the world, to fly to Cardiff. We would only expect the senior teams to travel for a face to face meeting if there was a particular reason, for example the first session of a new course or to manage a particularly challenging standardization process.

Once examinations have been taken the senior examiners will review examples of candidate work to ensure that the markscheme adequately covers the likely range of answers examiners are going to encounter. It is not possible for the markscheme to cover every possible answer, but the PE and senior team do need to ensure that it clearly explains the standard required for each mark.

It is important not to produce so detailed a markscheme or marking notes that they are not effective because examiners become confused or miss the most important points. The opposite is also true: a markscheme must contain enough information to ensure that all examiners give the same credit for particular answers. As a rough guide, if a specific comment is going to be relevant to less than 10% of candidates then it is unlikely to be useful in the guidance.

A key outcome of the standardization meeting is a series of **definitively marked scripts**. These are examples of candidates' work which have been marked by the PE, and which are then used to instruct and test other examiners in a process that is called the quality model. It is very important that these definitively marked scripts are correct as they are used to make decisions about the quality of other examiners. There are three uses for definitively marked scripts:

1. Practice scripts
2. Qualification scripts
3. Seed scripts

Practice scripts

- The purpose of practice scripts is to support examiners in learning the marking standard.
- Examples should show how to mark typical candidate's work or any common situations where examiners misunderstand the markscheme

These definitively marked examples of candidates' work are used by examiners to check their understanding of the markscheme and marking notes, and after reviewing these scripts, examiners should be confident that they can mark correctly.

When selecting practice scripts the PEs think about:

- supporting examiners in preparing to mark, so they would not include scripts that do not demonstrate a situation they are not likely to come across again
- how to explain why marks have been awarded using informative comments
- whether it is helpful and reasonable to exclude some parts of a candidate's script or put a note that they are not useful for examiners to learn from.

A good set of practice scripts would include (in order of priority):

1. examples that show how to mark typical candidate's work
2. any common situations where examiners misunderstand the markscheme
3. a good range from low to high marks will help examiners recognize where they might award marks and what a good (or bad) answer looks like
4. examples of exceptions and complex answers and how to deal with them.

Qualification scripts

- The purpose of qualification scripts is to demonstrate/prove examiners are marking to the correct marking standard.
- They comprise examples of the most common answers that candidates give access to the full mark range.

Before examiners are allowed to start marking "live" candidate work they must show us that they can mark to the correct standard. This is done by "testing" them with five examples of candidate's work which have already been marked by the PE so we know what mark they should be given.

If they do not give the same marks as the PE they are given feedback on what the correct mark was and why it was awarded. After they have reflected on this they have a second opportunity to show they now understand the marking standard.

A good set of qualification scripts would include (in order of priority):

1. examples of the most common answers candidates will give across the full range of marks (top, middle and bottom)
2. examples of work that require the examiner to have understood any common exceptions in the markscheme/common mistakes that candidates will make.

When selecting qualification scripts, PEs would not include scripts that are designed to "catch out" the examiners, but would select scripts that are difficult to mark but represent the kind of challenge that we expect examiners to deal with on a day-to-day basis. They would also look for examples of any important cases we need to be sure examiners will deal with correctly.

Seed scripts

- The purpose of seed scripts is to demonstrate/prove examiners are continuing to mark to the correct standard.
- Seed scripts would include common mistakes that examiners make and examples across the full range of marks

While we do not allow any examiner who has not understood the marking standard to start on live candidates' work, we also know that, over time, an examiner's standard can start to drift. For this reason, we periodically check their marking using seed scripts.

These seed scripts have already been marked by the PE so we know what mark should be awarded. These seed scripts look like every other piece of candidate work so an examiner is unaware that they are marking a seed.

If the examiner gives the principal's mark for a seed script, they continue with their marking. However, if they show they are no longer marking to the required standard, we intervene.

Initially, we provide feedback and guidance to allow the examiner an opportunity to re-establish the marking standard. They can then resume marking confident that all candidates will get a consistent mark whoever marks their work.

If an examiner is unable to re-establish the correct marking standard we would stop them marking any more candidates' work.

In general, an examiner will be asked to mark one seed in every ten live scripts they mark: however, not every tenth script is a seed, so it should not be obvious to them when they are marking a seed. We do vary seeding rates when appropriate.

A good set of seed scripts would include (in order of priority):

1. candidate work that represents common mistakes that examiners make
2. a good range of marks (top, middle and bottom).

Unlike qualification scripts, seed scripts are allocated randomly so every examiner will receive them in a different order.

When selecting seed scripts, PEs are very aware that at this point examiners are marking live candidate work and so the focus is on ensuring candidates are being awarded the correct marks. The means it is appropriate to include demanding or difficult scripts to mark. The purpose of these scripts is not to "catch the examiner out" but to make sure they are marking in the way we expect. It is also reasonable to use seeds to check that examiners are following all the marking rules, not just that their marking is correct. This means, rarely, they might include a seed that should be escalated to a team leader for some reason (for example, evidence of academic misconduct or a completely unexpected answer).

Team leaders should monitor and support examiners throughout marking. As soon as an examiner is stopped from marking because a qualification or seeding script has been marked incorrectly they should be contacted by their team leader to offer feedback and mentoring. Both the team leader and the examiner will be able to see the seeding script which was not marked to the standard and view the PE's marks and comments. The role of the team leader is to explain to the examiner why the marks they awarded were incorrect and they can together discuss the markscheme and its correct application. Examiners who continue to apply the markscheme incorrectly, despite the additional training provided by the team leader, will be stopped from marking completely.

Tolerances

- A tolerance reflects the legitimate differences in the marks awarded by different examiners to the same piece of work. Think about two teachers in your school marking a piece of work: both agree it is good, but one would award 46 and the other 47.
- The IB is committed to asking questions that test what is important, not just easy to mark.

For certain types of questions, where the answer is either right or wrong, we would expect an examiner to give exactly the same mark as the PE. For other types of questions, especially those which test understanding or analysis, there is a degree of judgment in the marks to be allocated. Two expert examiners with the same concept of what a good answer looks like would give closely related scores, but may differ by one or two marks.

We use the idea of tolerances to reflect these legitimate differences in opinion when reviewing examiner performance. For example, if the PE gave an essay a mark of 46 and the question had a tolerance of 2 marks, then we believe that any examiners who give it a mark between 44 and 48 have shown they are marking to the correct standard.

When reviewing questions with several parts or criteria, we monitor both the difference in overall mark (that is, total) and also differences for each part or criteria. An examiner must be within tolerance for both aspects to show they are marking to the correct standard.

Successful standard setting

If this stage of the assessment cycle has been successful we should have:

- consistency in marking
- a set of team leaders capable to intervene and help their examiners with almost any query relating to the marking of a component
- instructions, including the markscheme or marking notes, that clearly explain the rationale for marks awarded
- definitive marked scripts which are fit for purpose (and correctly marked).

Marking

- Successful marking is candidates being given a consistent and accurate score which reflects the quality of their work.
- The considered views of the PE are correct and every other examiner must reproduce this in their marking.
- Marks and grades are not the same thing—candidates may get more marks on an easier examination but should still receive the same grade.

What is marking—Consistent examiner judgment

The purpose of the markscheme is to support and remind examiners of the marking standard. The markscheme is not the definition of where marks should be given. The judgment of the PE is the definition of a "correct" mark. They then explain their thinking through the markschemes.

This is a very important principle because it means if the PE's considered opinion disagrees with the markscheme or marking notes, then it is the markscheme that is wrong and needs to be amended.

Thankfully, such cases are very rare but are extremely serious, because it means that other examiners who had relied on the markscheme would need to be re-briefed on what the correct marking standard should be. The IB would also need to re-mark all of the work that could have been affected by the error so that no candidates were marked incorrectly. PEs and their senior team take great care in standard setting to ensure the markscheme is accurate and informative to prevent such situations arising.

Examiners need to have passed qualifications and seeds before they are allowed to mark scripts. Interpreting the markscheme is not sufficient to be allowed to mark live candidate work. Similarly, teachers and parents should not assume that their interpretation of the markscheme is justification that a script has been marked incorrectly, although in such cases we would encourage the use of the "enquiries upon result" (EUR) service.

"Definitive marks"

The term "definitive marks" or "definitively marked script" means that the PE has made a formal judgment on the mark that should be awarded. These definitively marked scripts are then used in the quality model.

Marks and grades are not the same thing

An important aspect of carrying out, and using, summative assessments of candidates is to understand the difference between marking their work and grading their work.

- In marking, a candidate is given credit for the work they have produced against a markscheme or similar framework. This is an indication of the degree of the assessment task they got right. The mark itself has no other meaning.
- In deciding a grade, the examiner is making a judgment on the quality of the candidate's work against a defined standard which will take into account the difficulty of the task as well as the proportion of the task that was completed. The grade therefore has some meaning or relevance and is usually intended to be comparable with performances on other assessments.

Figure 10

Marking can be thought of as the quality (tastiness) of the food cooked, but the grade reflects the complexity of what they were trying to cook

It might be possible for a candidate to demonstrate a high grade from getting only a small proportion of a very difficult question correct, and be impossible to demonstrate the same grade by correctly answering many trivial questions.

As discussed in later sections, it is not necessary for the standard described by the grade to be explained by reference to what the candidate has attained, although this is the approach taken by the IB. There are other perfectly consistent and well respected systems where the standard is based on how the candidate performs relative to peers.

In our assessments, the IB generally uses marks as an indication of overall performance (compensation model) and then looks at how candidates with this number of marks performed to determine a boundary point (grade boundary) where students with more than that number of marks are awarded a particular grade. This process is explained in more detail in the "IB assessment practices" section.

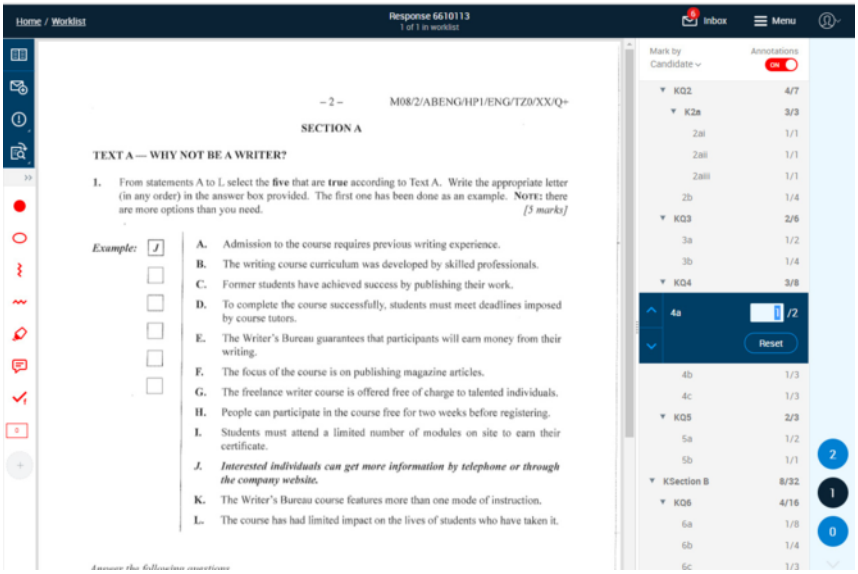
(Basic principles of) on-screen marking

- On-screen marking allows us to improve the quality of marking of our assessments.
- It allows the IB to monitor examiners' marking in "real time".
- It removes the time spent to post and return scripts and the risk of them getting lost.
- It allows the IB to "chop up" candidates' scripts and send different questions to expert markers (QIGs).
- Examiners can mark audio and video material and on-screen assessments as well as written scripts.

On-screen marking is where examiners are presented with images of the candidates' work electronically and are asked to mark it directly on the computer. They do this through a piece of software known as a "marking tool" which allows them to put ticks, marks and notes on the electronic copy of the script just like they would have done with a paper script.

Figure 45

Screen shot from RM Assessor, the e-marking tool currently used by the IB. © RM Results



As the marks are recorded directly by the computer this means we can monitor examiners as they mark. Having the scripts in an electronic format also has significant advantages.

First, it means that the scripts can be passed quickly to examiners as they need them to mark, removing the need to wait for paper scripts in the post and also removing the need to guess how many scripts an examiner will be able to mark—we now provide them with scripts as they need them. The benefit of having copies of the candidate's work in the system so that it cannot get lost or damaged is considerable.

Secondly, it allows us to introduce "test scripts" known as seeds into examiner marking so we can check they are maintaining the same standard throughout their marking.

Thirdly, it allows the IB to break up scripts into individual questions (known as QIGs) and ask examiners to focus on marking individual questions together rather than whole papers.

Finally, the use of on-screen marking supports the use of video and audio material in marking. This opens up a much broader range of candidate work that can be submitted as evidence, for example, recordings of performances or oral examinations.

Different types of markschemes

Analytic markschemes

Analytic markschemes are prepared for those examination questions that lead to a narrow range of expected answers from the candidates.

These markschemes can give specific instructions to examiners about how to break down the total mark available for a question, between different parts of the answer. Even with structured questions expecting highly specific answers, markschemes must provide examiners with sufficient information for them to mark consistently the main kinds of different approach that candidates might adopt and the common errors that they might make. It is inevitable that examiners will need to use their professional judgment in allocating marks to unexpected responses or alternative valid answers, but markschemes must provide as much guidance as possible in how to exercise that judgment.

Candidates often do not follow predictable patterns in what parts of a question they get right or wrong. This is an issue particularly with extended structured questions where a mistake in an early part could have an impact later on. While we design such questions so that if a candidate makes a mistake in the early part

of the question, the rest of that question does not become inaccessible to the candidate, even analytic markschemes need to provide explicit guidance on how to mark particular kinds of incorrect answer, and how to deal with following through candidates' working when they have made a mistake in part of a question.

This is a particular issue in subjects such as science and mathematics, where we expect examiners to award credit for partial success. Without structuring of an in-depth question, some candidates might not be able to achieve many marks because of a slight error early on in their response or because they have made a slight misunderstanding of the question and proceeded in quite the wrong direction. The most highly elaborated analytic markschemes are often found in mathematics.

Assessment criteria

Where an assessment task is so open-ended that the prospective variety of responses is too great to permit analytical markschemes to be meaningful, then descriptive (assessment) criteria are applied instead.

Assessment criteria do not generally refer to the specific content expected in a candidate's answer, although, where possible, they refer to the need for candidates to show certain kinds of content knowledge. The criteria concentrate more on the type of performance that candidates are expected to demonstrate, regardless of the specific details of the response. The range of different levels of expertise which could be demonstrated are reflected by level descriptors.

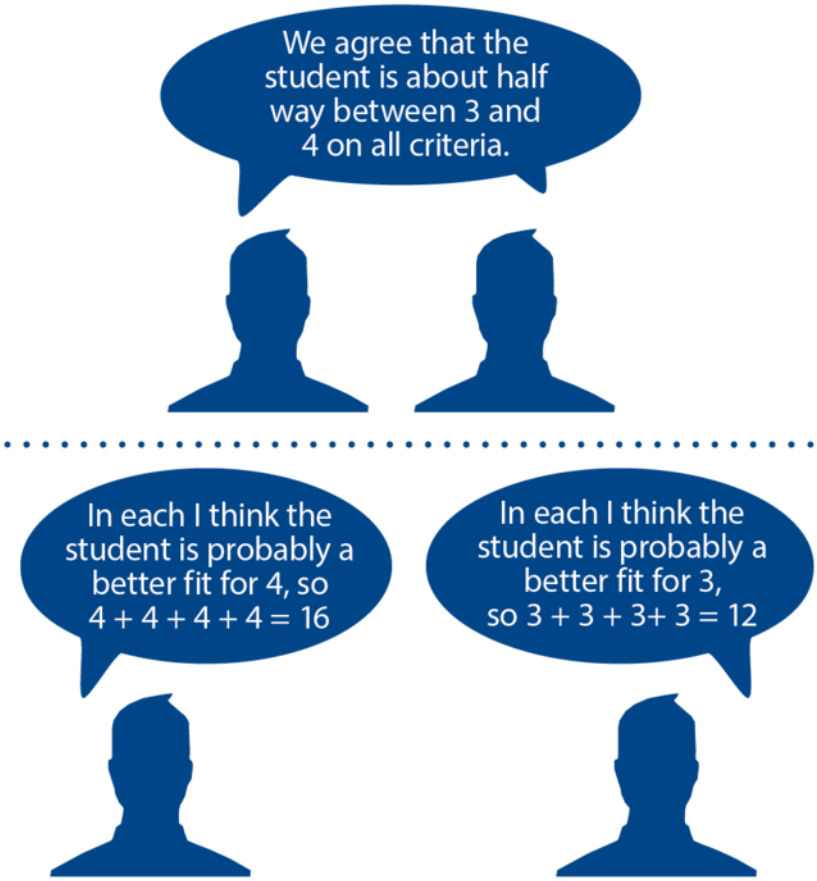
We use a "best fit" model in the application of criterion level descriptors. The examiner applying an assessment criterion must choose the achievement level that overall best matches the piece of work being marked. It is not necessary for every detailed aspect of an achievement level to be satisfied for that level to be awarded, and it is worth noting that the highest level of any given criterion does not represent perfection.

While it is perfectly possible to mark against a single criterion, it is more common to assess a single piece of work against several different criteria. The individual marks are then combined, with any criteria which are deemed more important being marked out of a higher total. It is very important that in such cases, the criteria are all independent of each other, it would not be fair for a candidate to gain credit in several different places for the same aspect of their answer.

The other problem with multiple criteria is that they tend to increase unreliability and inconsistency in marking. This is because the examiner is more likely to be faced with several "best fit" decisions where they could disagree by only one mark, but this can be multiplied if these differences all add in the same direction.

Figure 46

How assessment criteria can mean that examiners who broadly agree on the quality of candidate work can award very different marks



It is usual to have the same assessment criteria re-used every year, even though the questions asked of the candidates are different. This is because the underlying nature of what is being assessed remains unchanged, the different question generally relates to the specific details which are not explicitly set out in the criteria (unlike with analytic markschemes). The PE will often provide mark notes which provide guidance to examiners on how the assessment criteria should be applied to each question, and maybe examples of relevant details. When assessment criteria are used with internal assessment, both teachers and moderators should refer to the published teacher support materials, which give a number of examples of the application of the criteria.

An important point to keep in mind is that although criterion level descriptors are hierarchical in nature, and indeed often deal with the hierarchy of cognitive skills defined by Bloom et al (1956), there is no direct link between cognitive demand and criterion levels. Lower-level descriptors are not devoted only to the "simpler" cognitive skills, nor are higher-level descriptors reserved only for the "higher-order" cognitive skills. There should be a range of levels of achievement within each of the cognitive skill areas Bloom describes.

Holistic criteria—Markbands

Sometimes, it is not appropriate to separate out the different assessment criteria to mark a piece of work. This is usually where it is impossible to have distinct criteria which are independent of each other. In such cases, markbands are used instead of separate criteria. The markbands, in effect, represent a single holistic criterion applied to the piece of work, which is judged as a whole. Because of the requirement for a

reasonable mark range along which to differentiate candidate performance, each markband level descriptor will correspond to a range of marks.

Figure 47

An example of a markband taken from the 2015 DP history guide, paper 1 SL and HL

Marks	Level descriptor
0	The response does not reach a standard described by the descriptors below.
1–3	The response lacks focus on the question. References to the sources are made, but at this level these references are likely to consist of descriptions of the content of the sources rather than the sources being used as evidence to support the analysis. No own knowledge is demonstrated or, where it is demonstrated, it is inaccurate or irrelevant.
4–6	The response is generally focused on the question. References are made to the sources, and these references are used as evidence to support the analysis. Where own knowledge is demonstrated, this lacks relevance or accuracy. There is little or no attempt to synthesize own knowledge and source material.
7–9	The response is focused on the question. Clear references are made to the sources, and these references are used effectively as evidence to support the analysis. Accurate and relevant own knowledge is demonstrated. There is effective synthesis of own knowledge and source material.

The descriptors themselves tend to be quite lengthy, covering a range of potential qualities evident in candidates' work, and will relate directly back to the course objectives. As with assessment criteria, a "best fit" approach is used, with examiners additionally needing to make a judgment about which particular mark to award from the possible range for each level descriptor, according to how well the candidate's work fits that descriptor. For example, one markband level may cover the range 6 to 10 marks. The examiner will give a mark from that range according to how well the candidate's work fits the relevant level descriptor from the markband scale.

Research has shown that, where holistic (markband) and assessment criteria methods of marking have been applied to essay work that is amenable to both marking methods, there is little difference between the two in terms of reliability of marking (Wood 1991).

Additional support for examiners

Examiners, like all learners, benefit from having a range of ways of absorbing information.

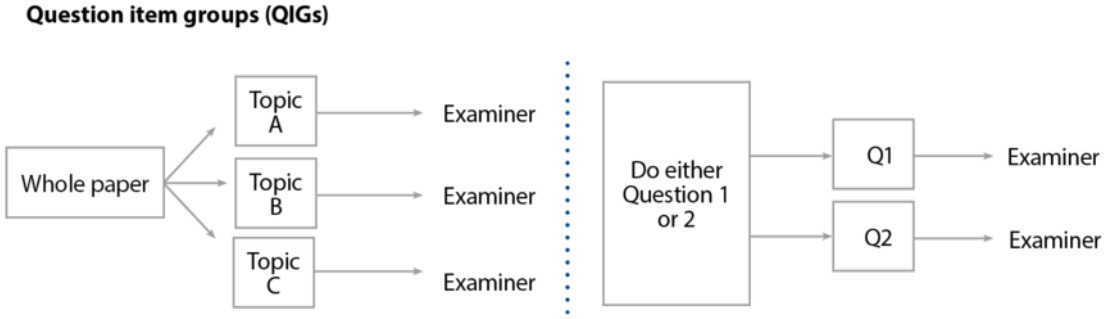
As well as written markschemes and assessment criteria, the IB is increasingly making use of screencasts to provide an audio-visual way of explaining the required marking standard. However, we also encourage examiners to ask questions and receive feedback from the senior team by telephone, messaging through the e-marking software and/or email.

This communication is built into our quality model. When examiners do not meet the required standard in a seed or qualification script they are required to obtain feedback from their team leaders. If their marking is of the required standard they are still given the opportunity for feedback.

Question item groups (QIG)

Figure 48

The two ways in which papers can be divided into QIGs



The idea behind QIGs is that it is easier for examiners to mark the same question many times rather than needing to mark all the questions on a whole script and then start again with question 1 for the next candidate.

There are two ways in which papers can be separated into QIGs. The first is to divide them by question or topic area. When doing this, it is important to keep related parts of questions in the same QIG, for example, if part (a) leads onto part (b) then both parts should be in the same QIG. This allows for examiners to give follow through marks and also to give credit when the candidate has put material that is relevant to the second part in their answer to the first part.

Conversely, the smaller the amount of detail in the markscheme the examiner needs to remember, the more consistent they are, so several small QIGs are better than one large QIG. This tension is taken into account when deciding how to divide up examination papers.

The second way that a paper can be broken up into a QIG is where there is a choice for the candidate on which question to answer. In this case, each optional question becomes a QIG, which may represent the candidate's whole script on that examination. For example, imagine a literature essay which allows the candidate to answer a question on poetry or on prose. All the poetry questions become QIG 1 and so an examiner who is an expert on poetry can mark all of these scripts. All the prose questions become QIG 2, and so an examiner who is an expert on prose can mark all of these scripts.

Each QIG will have its own practice, qualification and seed scripts, and the examiner will need to prove they can mark to the required standard on each QIG. While this may seem challenging, it means that an examiner who cannot grasp the standard on one particular question on a paper can still mark all the other questions rather than being stopped because they cannot mark the whole paper to the necessary standard.

Quality model

The IB now e-marks virtually all externally assessed work. Examiners' performance is monitored using an approach which requires examiners and the PE to mark the same scripts and then compares the examiners' marks with those of the PE for that component. The process works as follows:

1. The PE and a small number of their senior team of examiners mark a number of scripts during the week after the examination has taken place as part of the standardization process.
2. The definitive marks awarded by the senior examining team are recorded in the system together with their comments/annotations.
3. The IB, in consultation with the senior examining team, agrees how close to the definitive mark (the PE's mark) examiners need to be when they mark these scripts. This allowable difference is called the "tolerance".

4. All comments and marks are then hidden, and the scripts are put back into the pool for examiners to mark.
5. The definitively marked scripts are used for practice marking, the standardization process and seeds.

The concept behind the quality model is that examiners are able to learn throughout the marking process: when seeds are marked out of tolerance, examiners are immediately shown the definitive marks to allow them to understand how they have not applied the markscheme correctly and see what they need to do to improve. Examiners are also regularly shown the seeds they have marked within tolerance, so that they can aim to achieve the precise definitive mark as closely as possible.

Examiners should therefore regard seeds as opportunities for professional development. Most examiners will require feedback from seeds at some stage in their marking and the feedback should be valued rather than a cause for concern.

Practice scripts

- The purpose of practice scripts is to support examiners in learning the marking standard.
- Examples should show how to mark typical candidate's work or any common situations where examiners misunderstand the markscheme

These definitively marked examples of candidates' work are used by examiners to check their understanding of the markscheme and marking notes, and after reviewing these scripts, examiners should be confident that they can mark correctly.

When selecting practice scripts the PEs think about:

- supporting examiners in preparing to mark, so they would not include scripts that do not demonstrate a situation they are not likely to come across again
- how to explain why marks have been awarded using informative comments
- whether it is helpful and reasonable to exclude some parts of a candidate's script or put a note that they are not useful for examiners to learn from.

A good set of practice scripts would include (in order of priority):

1. Examples that show how to mark typical candidate's work.
2. Any common situations where examiners misunderstand the markscheme.
3. A good range from low to high marks will help examiners recognize where they might award marks and what a good (or bad) answer looks like.
4. Examples of exceptions and complex answers and how to deal with them.

Qualification scripts

- The purpose of qualification scripts is to demonstrate/prove examiners are marking to the correct marking standard.
- They comprise examples of the most common answers that candidates give access to the full mark range.

Before examiners are allowed to start marking "live" candidate work they must show us that they can mark to the correct standard. This is done by "testing" them with five examples of candidate's work which have already been marked by the PE so we know what mark they should be given.

If they do not give the same marks as the PE they are given feedback on what the correct mark was and why it was awarded. After they have reflected on this they have a second opportunity to show they now understand the marking standard.

A good set of qualification scripts would include (in order of priority):

1. Examples of the most common answers candidates will give across the full range of marks (top, middle and bottom).
2. Examples of work that require the examiner to have understood any common exceptions in the markscheme/common mistakes that candidates will make.

When selecting qualification scripts, PEs would not include scripts that are designed to "catch out" the examiners, but would select scripts that are difficult to mark but represent the kind of challenge that we expect examiners to deal with on a day-to-day basis. They would also look for examples of any important cases we need to be sure examiners will deal with correctly.

Seed scripts

- The purpose of seed scripts is to demonstrate/prove examiners are continuing to mark to the correct standard.
- Seed scripts would include common mistakes that examiners make and examples across the full range of marks

While we do not allow any examiner who has not understood the marking standard to start on live candidates' work, we also know that, over time, an examiner's standard can start to drift. For this reason, we periodically check their marking using seed scripts.

These seed scripts have already been marked by the PE so we know what mark should be awarded. These seed scripts look like every other piece of candidate work so an examiner is unaware that they are marking a seed.

If the examiner gives the principal's mark for a seed script, they continue with their marking. However, if they show they are no longer marking to the required standard, we intervene.

Initially, we provide feedback and guidance to allow the examiner an opportunity to re-establish the marking standard. They can then resume marking confident that all candidates will get a consistent mark whoever marks their work.

If an examiner is unable to re-establish the correct marking standard we would stop them marking any more candidates' work.

In general, an examiner will be asked to mark one seed in every ten live scripts they mark: however, not every tenth script is a seed, so it should not be obvious to them when they are marking a seed. We do vary seeding rates when appropriate.

A good set of seed scripts would include (in order of priority):

1. candidate work that represents common mistakes that examiners make
2. a good range of marks (top, middle and bottom).

Unlike qualification scripts, seed scripts are allocated randomly so every examiner will receive them in a different order.

When selecting seed scripts, PEs are very aware that at this point examiners are marking live candidate work and so the focus is on ensuring candidates are being awarded the correct marks. The means it is appropriate to include demanding or difficult scripts to mark. The purpose of these scripts is not to "catch the examiner out" but to make sure they are marking in the way we expect. It is also reasonable to use seeds to check that examiners are following all the marking rules, not just that their marking is correct. This means, rarely, they might include a seed that should be escalated to a team leader for some reason (for example, evidence of academic misconduct or a completely unexpected answer).

Team leaders should monitor and support examiners throughout marking. As soon as an examiner is stopped from marking because a qualification or seeding script has been marked incorrectly they should be contacted by their team leader to offer feedback and mentoring. Both the team leader and the examiner will be able to see the seeding script which was not marked to the standard and view the PE's marks and comments. The role of the team leader is to explain to the examiner why the marks they awarded were incorrect and they can together discuss the markscheme and its correct application. Examiners who

continue to apply the markscheme incorrectly, despite the additional training provided by the team leader, will be stopped from marking completely.

Tolerances

- A tolerance reflects the legitimate differences in the marks awarded by different examiners to the same piece of work. Think about two teachers in your school marking a piece of work: both agree it is good, but one would award 46 and the other 47.
- The IB is committed to asking questions that test what is important, not just easy to mark.

For certain types of questions, where the answer is either right or wrong, we would expect an examiner to give exactly the same mark as the PE. For other types of questions, especially those which test understanding or analysis, there is a degree of judgment in the marks to be allocated. Two expert examiners with the same concept of what a good answer looks like would give closely related scores, but may differ by one or two marks.

We use the idea of tolerances to reflect these legitimate differences in opinion when reviewing examiner performance. For example, if the PE gave an essay a mark of 46 and the question had a tolerance of 2 marks, then we believe that any examiners who give it a mark between 44 and 48 have shown they are marking to the correct standard.

When reviewing questions with several parts or criteria, we monitor both the difference in overall mark (that is, total) and also differences for each part or criteria. An examiner must be within tolerance for both aspects to show they are marking to the correct standard.

Challenging and unusual scripts

If examiners come across scripts they cannot decide how to mark, their first recourse should be to contact their team leaders to ask for advice. If they still do not believe they can mark it fairly, then they can send it to an examiner more senior than themselves to mark. Particularly problematic scripts will then come to the attention of the PE who can provide the definitive decision on how to mark the script.

In a similar way, any unusual scripts or candidates who have answered modified question papers will be marked by the PE, who will need to carefully balance the marking to ensure it is equivalent to the same standard as applied to other candidates.

This marking takes place after grade boundaries have been determined, so that senior examiners have a common understanding of the grades in the context of this particular assessment session. While a mark is being awarded, in these particular cases we require our examiners to consider the context of the grade boundaries to ensure that the unique marking standards set for these candidates are appropriate with the grade boundaries that have been determined with reference to the main cohort.

School connections

The IB's principle is that examiners should not mark their own candidates' work.

To manage this, we instruct all examiners to inform us of any connections they have to candidates and schools so we can ensure that they do not receive their work to mark. We also have a specific conflict of interest policy for the Assessment Division in IB to cover people working for the IB who have links with schools.

If you have any suspicions concerning examiners or IB staff not declaring a school or candidate connection, please contact complaints@ibo.org or IB Answers.

Examiner comments

> One cannot manage too many affairs: like pumpkins in the water, one pops up while you try to hold down the other.

(Chinese proverb)

The purpose of IB summative assessment is to measure a candidate's performance and we require examiners to focus solely on marking candidates' work to the required standard. We only ask examiners to make comments when it helps them in doing their marking.

Writing formative feedback for either candidates or teachers requires the examiner to determine the correct mark for a piece of work and then try to explain how the work could have been improved. We reflect that the second task is at the heart of what good teaching means, and is not a trivial task. It requires time and thought which will draw the examiner away from their core task of marking to a consistent standard. In simple terms, we want them to do one task (marking) to a high standard, not two tasks (marking and feedback) to a lower standard.

The examiner can also only make their judgment on the one piece of work they have available, and experienced teachers will draw upon a wide range of information when deciding how to offer feedback to a candidate. This means the quality of any examiner feedback will suffer from having less insight than that of the teacher.

For all these reasons, the IB is very clear to its examiners that they should only mark the candidate's work according to the correct standard and not add comments to provide feedback to the candidate or teacher.

Examiners are required to indicate clearly where marks have been awarded, and, if there could be ambiguity, to clarify with appropriate comments. This supports the IB in checking standards and also provides transparency for schools on where marks have been awarded.

The exception to this principle of only commenting where it supports the marking is where a school has requested a category 1 EUR report. In this case, a senior examiner will address the specific concerns that a school has raised when requesting the EUR, which will go beyond the usual level of detail we expect of examiners.

Aggregation

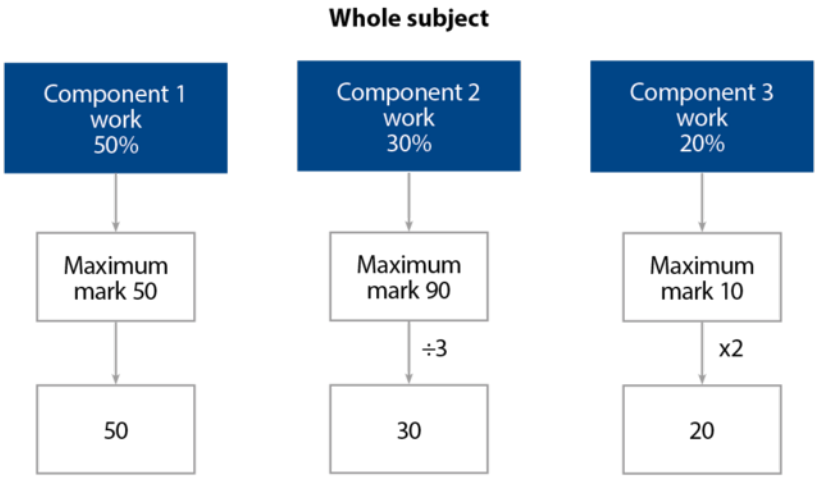
- Aggregation is the process of combining components to generate an overall result.
- In order for each component to contribute in the correct proportion to the final mark (weighting) it may be necessary to scale the component marks.
- The IB uses a "compensation model" where candidates can offset poor performance in one question or component with high attainment in another.
- A candidate's final subject grade is determined from the aggregation of component marks, and not from component grades.

Aggregation is the process of combining marks (and boundaries) from the different components to form a final mark or overall grade boundary. To achieve this, overall component marks (or boundaries) may need to be scaled.

Scaling is carried out to preserve the desired weighting for each component in terms of its contribution to the overall assessment for the course. It means multiplying or dividing component marks so that they contribute correctly to the overall total for the subject. The same applies to the grade boundaries set for the component, which would have to be determined initially out of the maximum marks for the components.

Figure 59

Example of scaling in a subject with three components



The concept of weighting is to reflect the relative importance that the IB places on the elements being assessed in contributing to the final outcome. For example, if a component primarily tests interpretation of data or sources, and has a weighting of 30%, then this implies that, compared with the other objectives of the course, interpretation represents about 30% of what is important. Often several components will test similar objectives and so this calculation is less meaningful.

A secondary, but important, aspect of weighting is that it allows us to set the total number of marks in an assessment that is appropriate to the tasks and the marking criteria rather than trying to force it artificially into an overall total.

It is a very important point that we do not require an individual candidate's marks to match the component weighting. We recognize that different candidates have different strengths, and this is why we use a variety of assessment instruments. It is important that we review the actual weightings of the entire cohort's results against those set out in the design of the subject. If one component's contribution is much higher in a session than intended, then this may indicate that the paper was particularly easy and we should look at other evidence in grade award. Alternatively, if one component's contribution is much higher than intended on a regular basis, then we need to review the design as the assessment is not working the way it was intended.

The approach outlined may not reflect the more sophisticated methods of weighting, combining (aggregating) and scaling described by, for example, Wood (1991: Chapter 10), but is based on sound criterion-related principles and supports the IB's principles of assessment.

This approach to aggregating the final mark means that a candidate can offset poor performance in one component with high performance in another, as it is only the total mark which impacts on the final grade. This concept is referred to as a "compensation model" to contrast it with a "mastery model" where a candidate would need to show the required level in all components to be awarded that grade.

It is worth stressing that a candidate's final subject grade is determined from the aggregation of component marks, and not from component grades. Because each component grade represents a range of marks, it is quite possible for two candidates with the same component grades to be awarded different subject grades.

Moderation

- Moderation is the checking of teachers' marking standard, it is not about re-marking candidate's work.
- Successful moderation means that candidates would receive the same internal assessment mark even if they had gone to another school on the other side of the world. We call this the global standard.
- The evidence used in moderation is the teachers' explanation of why they have awarded marks, not just the quality of the candidate's work.

Use of dynamic sampling—the new approach to moderation—means that the IB is confident that every moderator is matching the PE's standard, so that only one stage of moderation is required.

What is moderation?

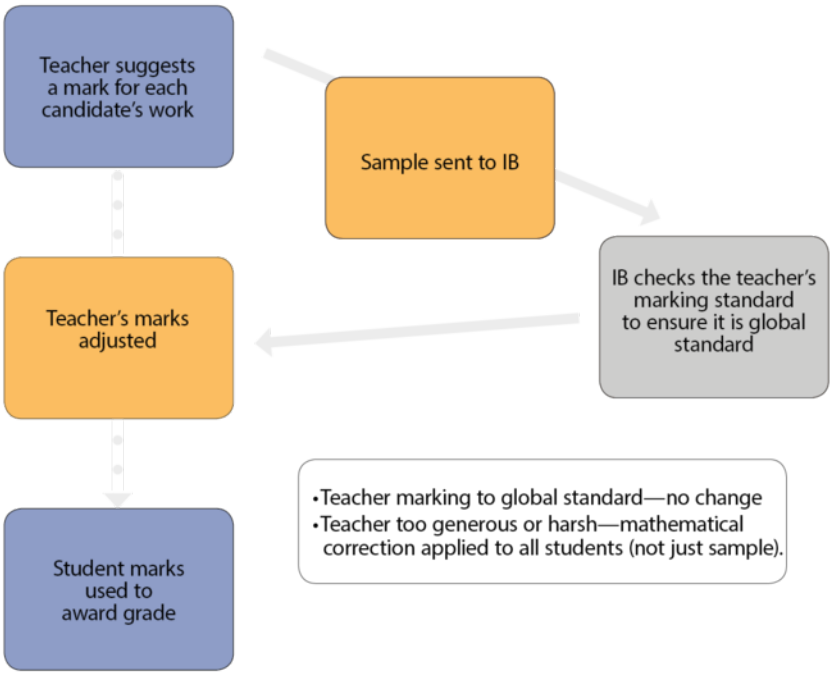
In many cases, the trait we wish to assess in a candidate cannot be tested in a formal, time-limited examination. In such cases, the most valid approach is to ask the teacher to carry out the testing with an internal assessment. The section on "The role of classroom-based assessment and internal assessment" sets out why this is an appropriate approach.

While this produces meaningful results, it also creates a risk that different teachers have different interpretations of what the marking standard is. So, two teachers in two different schools could award the same piece of work two different marks. The IB has a comprehensive process for training and testing examiners so that they all have a common understanding of the required standard, but this is not feasible for all teachers.

Instead, we have confidence that our teachers are marking to a consistent standard for all their candidates so that we only need to apply an arithmetical factor to them if necessary in order to bring them in line with the global standard. We do this by asking for a sample of their marking and comparing this to the PE's standard. From the data this comparison provides, where necessary, we calculate a mathematical formula which adjusts all of the marks provided by each teacher.

Figure 49

Overview of moderation



The diagram above explains how the sample is moderated by the examiner. Based on a statistical comparison between the two sets of marks, if required an adjustment is made to the teacher's marks for all candidates at the school (for that component). If the teacher is consistently under- or over-marking, this adjustment will be the same for each of the teacher's marks, but if the teacher is under or over-marking either at the top or bottom of the mark range, this adjustment may vary across the range of the teacher's mark.

Important points to remember

- With moderation, the aim is to check how accurately and consistently the teacher has applied the assessment criteria in his or her marking of the candidates' work.
- As a result of moderation, a school's marks may be lowered, raised or remain the same.
- A moderation factor does not mean that the teacher's marking is of poor quality, it only means that it is more or less generous than the global standard.

For practical reasons, the IB moderates schools rather than individual teachers, and so it is very important that all teachers in a school ensure they are marking to the same standard so that the IB is fair when it applies one moderation factor to all teacher marks.

Figure 50

Considerations for a coordinator when reviewing internal assessment marks that will be moderated

Are subject teachers* in the school **all** marking to the same standard?

Are teachers are consistent in marking all candidates to the same standard?

Are teachers providing clear explanations on why they awarded a particular mark?

* Subject teachers who are responsible for providing internal assessment marks

For further details of the technical aspects of the calculation of moderation factors see "Annex 1: Moderation of internal assessment".

Selection of candidate work

- All candidates should be given a fair mark and so any candidate is appropriate to be used as an example of the teacher's standard for moderation.
- Two teachers can have different expectations of very good or very poor work so it is important that sample material covers the full range of marks.
- To ensure transparency and fairness between all schools, the IB needs to be able to see evidence of all work that is contributing to a candidate's final mark.

In order to ensure transparency and remove any perception of academic misconduct, it is important that the IB, rather than the school, identifies the work that is to be sampled for moderation. The IB uses broad guidelines (see below) to ensure that the moderation factor is as reliable as possible and, within the limitations of these guidelines, that the actual selection of candidates is random.

The first principle is that the IB should use the smallest possible sample to obtain a reliable moderation factor for any particular school. This minimizes burden on the school and also the costs to the IB which would then be passed on to schools through the exam fees. It is reasonable for the number of pieces of candidates' work in a sample to be different for different schools if the IB requires more examples of teacher marking to determine a robust moderation factor. This is why we sometimes need to request additional samples. As we may need to request additional sample work during the process of moderation, all candidates' work must be available until the issue of results.

The second principle is that we must be confident that our moderation factor is fair to all candidates across the range. From experience we know that teachers can have different expectations at different points of the mark range and that a teacher who is more generous than the global standard for poor quality work may be harsher than the global standard for high quality work. For this reason, the IA sample is carefully selected to ensure that the entire mark range of the school is appropriately represented.

We also tend not to select candidates for the moderation sample who have attained full marks. This is to be fair since it allows candidates in the higher mark range the possibility to be moderated upwards if a teacher is too harsh in their marking. We also would not usually select work from a candidate who had been awarded zero marks.

Unusual and atypical work

It is important that every candidate is marked fairly by the teacher: therefore for every candidate we would expect the teacher's marks and comments to be representative of their standard. This means that, in nearly every case, we would expect a school to submit work to the IB if requested for moderation.

Moderation is designed to check that teachers are marking to a global standard, even if a piece of work is particularly challenging to mark. In such situations the teacher may well need to give a more detailed justification of why they have awarded that particular mark, for example they provided additional help to the candidate, they would explain this in their marking and the examiner would take this into account when reviewing the teacher's marking of the candidate's work.

Failing to find a moderation factor

In some cases, it may not be possible to calculate a moderation adjustment using the submitted sample work. This happens if the difference between the examiner and the teacher is inconsistent, or the examiner believes the teacher is being too generous, or maybe the teacher is being too harsh compared with the global standard. In these cases, we will request further work from the school so we can be sure that a fair moderation adjustment can be applied. For this reason, all candidates' work must be available until the issue of results.

This situation sometimes occurs because several teachers have marked the work and they had not successfully established a common standard. As we apply a single moderation factor for each internal assessment per school, it is important that teachers moderate themselves before the work is submitted.

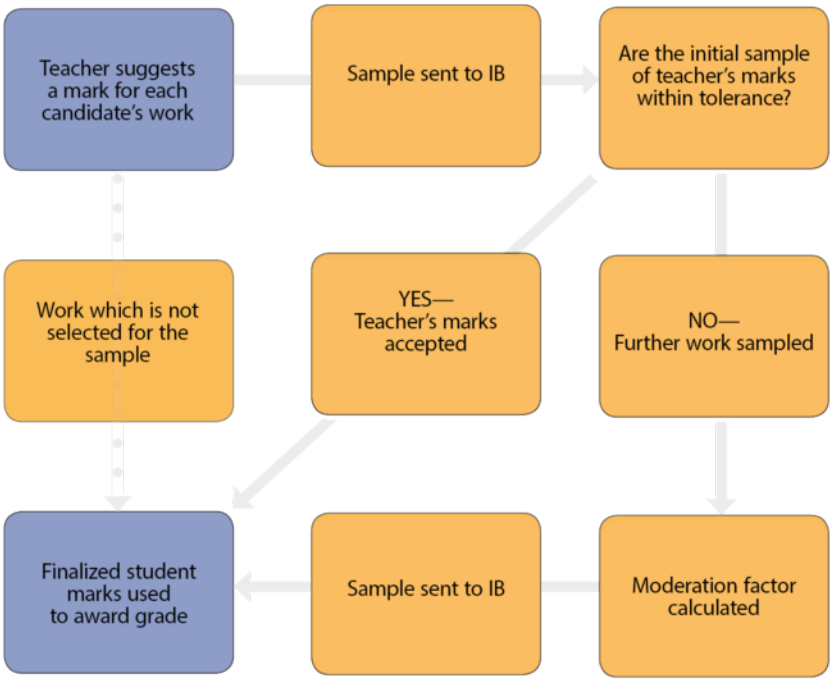
Dynamic sampling

- Dynamic sampling moderation means that if a teacher is marking within tolerance we accept his or her marks.
- If a teacher's mark is outside of tolerance, we apply an arithmetical moderation factor.
- Every examiner who is checking teacher IA marking is checked by a quality model similar to that used in examinations.

The principle of moderation using dynamic sampling is the same as we use for our examiners. If a teacher demonstrates that they understand the global standard through being within tolerance on their sample, we accept all their marks are to the required standard. If they are outside tolerance, then we calculate and apply a moderation factor. This is explained in the diagram below.

Figure 51

Flow chart for dynamic sampling moderation



The differences that schools will experience with dynamic sampling moderation, as opposed to the previous approach, is that they are far less likely to have a small moderation factor applied to their marking because, in such cases, they will be judged to be within tolerance. Another change is that the IB will be able to provide more detailed feedback since only one examiner's views will be required as a result of the quality model below.

For moderation to be fair, we need our examiners to understand the global standard set by the PE and to moderate to it. We use a similar system to the examiner quality model to ensure this happens.

Examiners who will be undertaking moderation receive training in the global standard and are then checked to ensure they are moderating to this standard. As described in the "Quality model" section the PE prepares three types of definitively moderated scripts.

- Practice scripts—to explain the global standard
- Qualification scripts—to check examiners understand the global standard
- Seed scripts—to check examiners are maintaining the global standard

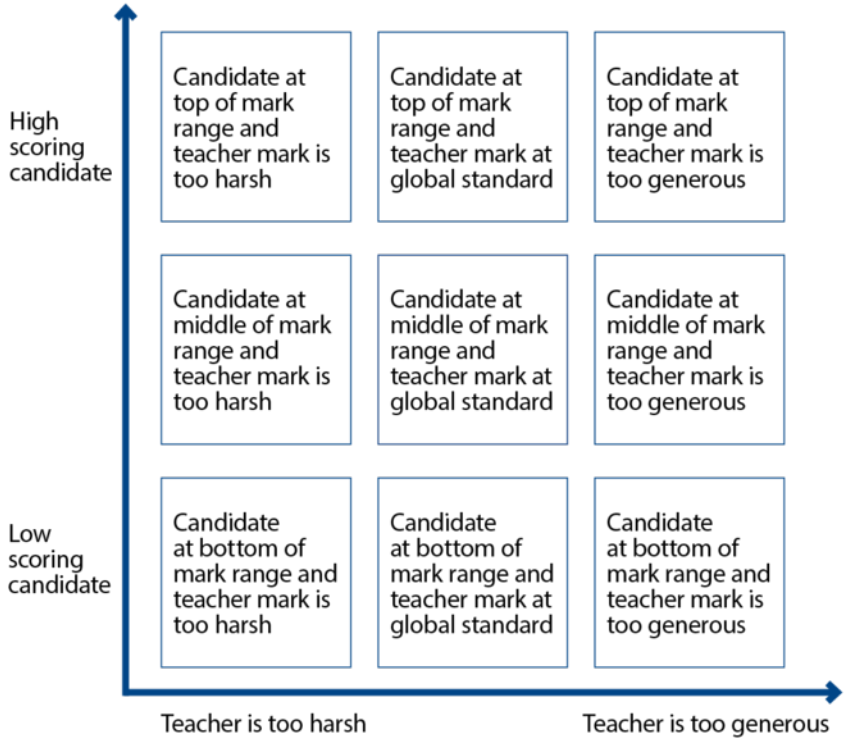
Unlike the previous moderation approach, with dynamic sampling, examiners are presented with work from a range of schools in no particular order. This mitigates against examiners forming an opinion on whether a teacher has over- or under-marked from the first candidate they see, and then looking for this pattern in the rest of a school's sample. It also means we can include seed scripts without it being obvious that it is a seed.

The same examiner will still review all the scripts from a single school, but will no longer review them together, and we ask them to provide summary comments on the schools marking once it has been determined whether a moderation factor is required.

The quality model in dynamic sampling is more complex than with other marking. Unlike with examination papers when an examiner only needs to be checking across the entire mark range, when preparing for moderation we also need to check that they are confident with the teacher's marks which may be too harsh, too generous or correct. The diagram below shows this.

Figure 52

Range of definitively moderated scripts required



Previous system—The moderation hierarchy

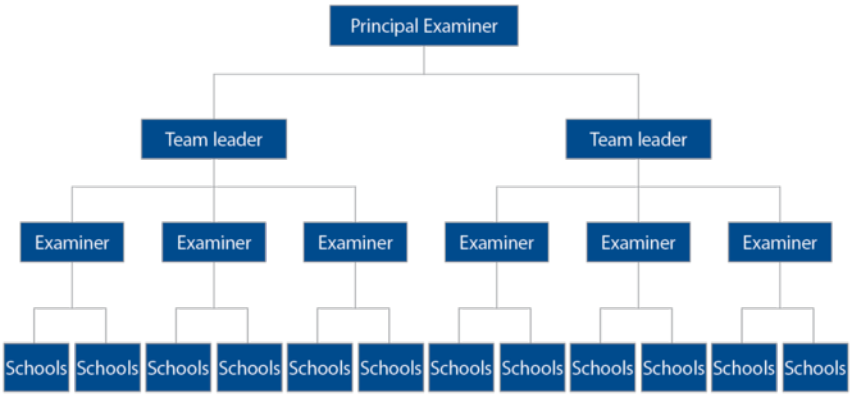
- The moderation hierarchy system is being replaced by dynamic sampling whereby we ensure that all examiners are aligned with the global standard before they moderate.
- The previous system meant that each examiner (below the PE) also had a moderation factor applied to their decisions. The final moderation factor applied to a teacher's marks was the combination of all these factors so that each set of teacher's marks was aligned with the global standard set by the PE.

Before the introduction of dynamic sampling, the IB used a second moderation approach to check that its examiners were reviewing teachers' marks correctly. This meant that the PE would review a sample of the decisions made by their senior team and an adjustment would be made to their decisions to make sure they were in line with the PE's standard.

This created a hierarchical process which ensured that the final marks awarded to every school are in line with the standard of marking set by the PE. The different levels of the hierarchy for a typical large entry IA component is illustrated below:

Figure 53

Moderation hierarchy



The diagram above shows that:

- A school's marks may be adjusted based on the sample submitted to an examiner.
- Every examiner's marking is also reviewed and adjusted based on a sample of their marking, which is submitted to a senior examiner ("team leader").
- In turn, team leaders' marks may be adjusted based on a sample of their marking which is submitted to the principal examiner.

Therefore, there is a chain of moderation where a series of adjustments can be made to a school's marks before a final, moderated mark (aligned to the principal's standard) is awarded.

Internal assessment (IA) feedback

The purpose of IA moderation is to ensure that all teachers are marking to the same standard, and ultimately, the IB would like no moderation factors to be applied to any school's work. To help teachers understand how they are varying from the global standard, the IB provides feedback on the IA marking so that teachers can understand why a moderation factor has been applied.

The feedback provided to schools is not intended to explain how the candidates in the sample could have achieved a better result.

Grade awarding (and aggregation)

- What is grade awarding?
- Grades should mean the same whichever session a candidate takes their exam in.
- The grade award process decides how to convert between marks and grades to ensure that this is the case.
- Grade boundaries are determined using a range of evidence including both expert judgment of candidate work and cohort results.
- It is the overall grade that a candidate receives that is the most important aspect, not any individual components.

It is important to remember that marks and grades are not the same thing. For more details on why this is the case see the section titled "Marks and grades are not the same thing".The grade award process is how the decision is reached on how to convert between marks and grades.

This decision is made through several days of discussion between the PE, CE, senior examiners and the IB. It draws upon a range of evidence (described in more detail in "Evidence used in grade award") to reach a conclusion that is fair to candidates this year, but also to candidates who have taken the subject in previous years. Finally, recommendations are made by the CE to the IB on what the outcomes for this examination session should be.

Through the process of grade awarding, evidence is also gathered on how effective the assessments have been. This intelligence is then used to improve future sessions. In particular, in the first year of a new course, the IB curriculum manager, who has responsibility for developing the course curriculum, will support the grade award meeting to help align the assessment outcomes with the intention of the course aims and to inform future curriculum development.

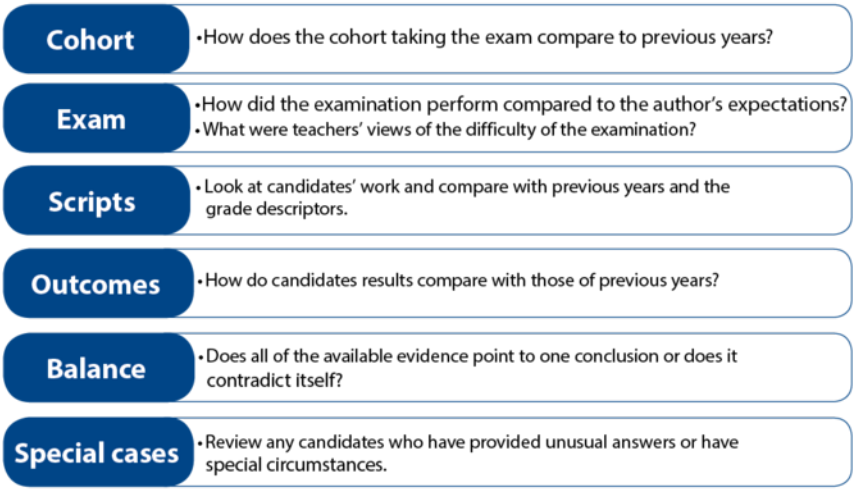
The main outcome of the discussion is the "grade boundaries", the minimum number of marks that are required by a candidate to obtain each grade. As it is not practical to make a detailed judgment for every single grade in every subject, the IB asks its examiners to make a recommendation on several "judgmental grade boundaries" and calculates the remaining boundaries arithmetically. In both MYP and DP/CP, where there are seven grades, the judgmental grade boundaries are the 2/3 boundary the 3/4 boundary and the 6/7 boundary, as these represent the highest level of understanding and knowledge of a subject (6/7), the performance of a candidate with basic knowledge and understanding (2/3) and the level of secure understanding and knowledge (3/4).

Grade boundaries should change every session. This is because the questions that the candidates are asked to complete will be different and so grade boundaries need to vary to reflect easier or more difficult tasks. While the IB makes every reasonable effort to ensure that its examination papers are of the same level of difficulty every year, because of their high stakes nature we do not trial examinations before they are sat because of the risk of the questions being made public.

In making the grade award decisions examiners and the IB consider the following aspects:

Figure 54

Aspects that need to be considered during grade award process



In many circumstances, the IB has several assessment components which are combined to give an overall grade. Details of how this is done is given in the aggregation section below. The most important aspect of this that examiners need to keep in mind during the grade award is that it is the overall result which is most important. If necessary, individual components can be less perfect to achieve a fair overall outcome.

Formally, the purpose of the grade award process is:

- to establish the point(s) in the distribution of candidate work where there is a change in which grade descriptor best describes its quality.
- For each component, determine the marks which are the grade boundaries.
- To ensure that the combination of these grade boundaries at the course level represent a fair awarding of grades.
- to consider which grade to award to any atypical or unusual candidate responses.

A grade award process has been successful if:

- script judgment and outcome evidence are in broad agreement; taking cohort information into account
- any variation in school performance can be explained
- the CE and IB Chief Assessment Officer are confident that assessment standards have been maintained.

Judgmental and interpolated grade boundaries

The judgmental grade boundaries are those which are recommended by the CE based on discussions during the grade award process. For the DP, CP and MYP, these judgmental boundaries are the 6/7 boundary, the 2/3 boundary and the 3/4 boundary as these represent the highest level of understanding and knowledge of a subject (6/7), the performance of a candidate with basic knowledge and understanding (2/3) and the level of secure understanding and knowledge (3/4).

The remaining boundaries, 1/2, 4/5 and 5/6, are calculated arithmetically based on the judgmental boundaries and so are known as the interpolated boundaries.

While these interpolated boundaries are based on evidence from consideration of the candidate's work, they are an important part of reviewing overall candidate attainment. If there was a significant shift in the proportion of candidates receiving a grade 5 or a grade 1, then this needs to be discussed in the context of

the cohort and may lead to either a reconsideration of the judgmental boundaries, or in exceptional circumstances, a review of candidate work at these boundaries.

Impact of eAssessment on grade award

A grade award process needs to take place regardless of the form of assessment that takes place. The introduction of eAssessment will make no difference to the principles of grade award.

Where eAssessment may have an impact on the grading process is by allowing assessments to test the aims of the course more effectively. Often it is difficult for examiners to see evidence of analysis, investigation or collaborative thinking in paper examinations, meaning that these aspects of the grade descriptors are under-represented; eAssessment may help address this.

Evidence used in grade award

Grade awarding is an evidence-based process that needs to draw upon a range of information including:

- teacher feedback on the assessment
- expert judgment from examiners on examples of candidates' work
- review of statistical information comparing this year's candidates' outcomes with previous years.

No one type of evidence is more important than any other: all must be balanced equally in coming to a conclusion.

Considering this year's cohort

The first task of the grade awarding process is to consider how similar the cohort taking the assessment is to previous years. If many of the same schools are taking the assessment this year compared with previous years, then any differences in performance are likely to be due to the difficulty of the examination papers. If, on the other hand, there is a large number of new schools taking the subject for the first time, or a high proportion of candidates are resitting the assessment, then we might decide any differences in performance need to be reflected in grade outcomes.

Examples of the kind of factors that will be considered by examiners when comparing the cohorts taking the assessment are:

- changes in the number of candidates taking the assessments
- changes in the proportion of candidates taking the assessments in English, Spanish and French (and other languages where appropriate)
- number of new schools and number of students in those new schools
- any changes in the options taken by students.

Feedback on the assessment

The next task will be to consider how the assessment performed relative to expectations. If a particular question proved far more challenging than expected, or items expected to discriminate between grade 6 and grade 7 candidates failed to do so, then senior examiners will need to take this into account when determining grade boundaries.

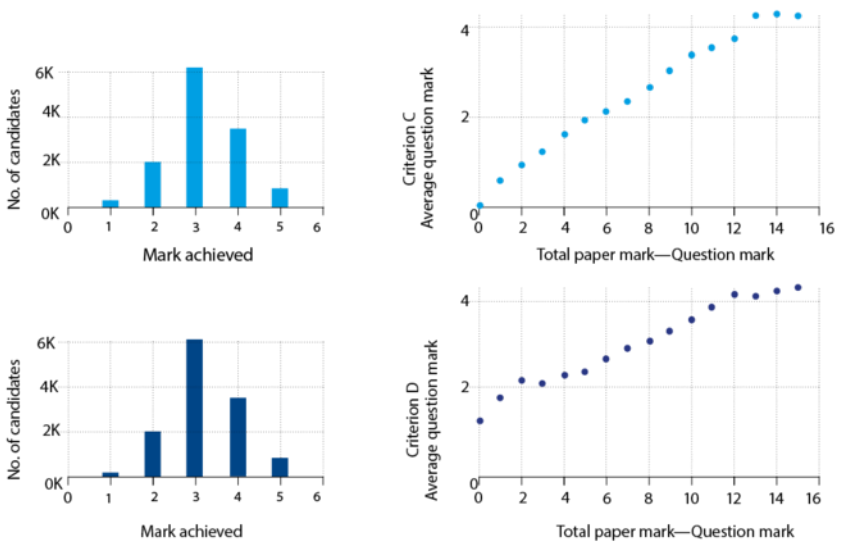
Senior examiners will base this discussion on teacher feedback on the examination papers (primarily the IB's teacher feedback (G2) forms) and statistics on the performance on individual questions/items (see diagram below). The team will also have reports from individual examiners which will have been summarized by examiners.

Figure 55

Example of the item level statistics available to examiners

Item summary statistics						
Question	Question	Total	Proportion of candidates attempted	Average question item mark	Highest mark achieved for item	Population standard
01 Criterion A	12,071	42,042	28.71%	2.7	5	0.95
01 Criterion B	12,071	42,042	28.71%	2.6	5	0.94
01 Criterion C	12,071	42,042	28.71%	2.8	5	0.84
01 Criterion D	12,071	42,042	28.71%	3.2	5	0.79
02 Criterion A	29,848	42,042	71.00%	2.8	5	0.98
02 Criterion B	29,848	42,042	71.00%	2.7	5	0.93
02 Criterion C	29,848	42,042	71.00%	2.9	5	0.83
02 Criterion D	29,848	42,042	71.00%	3.2	5	0.79

Correlation between Item and rest of component	
Question	Correlation coefficient
01 Criterion A	0.8044
01 Criterion B	0.8018
01 Criterion C	0.7983
01 Criterion D	0.7253
02 Criterion A	0.7954
02 Criterion B	0.7944
02 Criterion C	0.7865
02 Criterion D	0.7320



The team will also draw upon their own experiences of marking candidates' work during the session. This is why it is important that the grade award discussions involve a cross-section of those who have marked each of the components during the session. As the chairperson for these discussions, it is not essential that the CE has marked candidate work, indeed it can be helpful to have no preconceptions, but this means there is additional responsibility of the individual PEs to provide the necessary insight into the general quality of candidates' work.

Reviewing script evidence

The grade descriptors set out the characteristics that we expect to see in candidate work for each of the grades. During this phase of the grade award process, it is essential that examiners focus not on the marks awarded but on the nature of actual candidate responses and how well these match the grade descriptors. For this reason, as much as possible, scripts are selected that have a generally even level of response across the whole paper, rather than scripts that have scored highly on some questions and poorly on others.

Before starting the script review it is often helpful to look back at examples of work from previous years to remind examiners of the expectations for the different grades. These boundary scripts must be available during grade award process.

Before the grade award meeting, the senior examining team, and in particular the PEs, will submit provisional grade boundaries for each component, indicating at which marks they feel the boundaries should lie, based on their past experience of the expected standards. These provisional boundaries, together with the consensus of how each paper has functioned and an awareness of the overall distribution of marks, allow the IB subject manager to suggest a range of marks (sample scripts) which the senior team need to start their review of candidate work for each grade.

Each senior examiner should review the selected candidate script and determine which grade descriptor best reflects the quality of work. It is important to focus on the script as a whole and not be influenced by the marks awarded. It is helpful to know which questions were designed to provide the evidence of the higher grade, but these should only form part of the overall decision. This is a challenging task as even candidates who have a relatively even level of response will often show qualities of a wide range of grades across their answers, and examiners will need to judge which grade is the best fit. It is acceptable and often helpful to indicate whether a particular script is just within a grade descriptor or almost in the next higher grade (often indicated, for example, by a 7- or a 6+).

It is very important that examiners minimize any possible preconceptions or bias when undertaking this task, and so it is important that they do not discuss their views with fellow examiners until everyone has recorded their independent result. Similarly, it is helpful not to inform the examiners (excluding the CE) of the findings of the statistical analysis until they have completed this stage.

Determining which grade is the best fit for a script is a very subjective exercise, and it is likely there will be variation between the different examiners. Marks and grades also represent subtly different performance, and so it is not unreasonable for a candidate, who has shown good understanding but made a number of mistakes, to fit a higher grade than a candidate who has excelled at the easiest task (and so gained many marks) but not shown the same depth of understanding, despite the latter candidate having slightly more marks. It is also important to remember that this script review only considers a relatively small number of examples of candidates' work and any decision needs to reflect this.

Research has shown that examiners are most skilled at observing when scripts are of a different quality to others they are looking at (see "Alternative forms of marking"), so the IB asks examiners to start at the highest mark in the sample and to work down until they believe they have reached the point that they have stopped consistently seeing evidence of the higher grade. Then they should start at the lowest mark in the sample and work up until they start consistently seeing evidence of the higher grade. If necessary, the range of marks in the sample can be increased.

When all of the senior examiner grading decisions are collected together it should indicate the range of marks where a grade boundary should lie. This is called the "zone of uncertainty" and represents the lowest mark where reasonable evidence exists to place the grade boundary to the highest mark. There is not a precise definition of how to set the zone of uncertainty, it should be agreed through a discussion of the senior examiners based on their individual decisions.

Figure 56

Example of script review outcomes. In a real grade award more scripts at each mark would be considered.

Script no	Mark	Examiner 1 (PE)	Examiner 2	Examiner 3	Examiner 4
1	51	4	4	4	4
2	51	4	4	4	4
3	50	3+	4	4	3+
4	50	4-	4	4	3+
5	49	4-	4-	4-	4-
6	49	3+	3	4-	3+
7	48	4-	3+	4-	3+
8	48	3+	4-	4-	3+
9	47	3	3	3+	3
10	47	3	3	3	4-

Zone of uncertainty

The only exception to this process is for multiple-choice question papers. Experience shows that making judgments about grade boundaries based on the quality of candidate work is very difficult for papers made up only of multiple-choice questions. This may be because the responses contain very little evidence of what candidates have actually done, on which to make a judgment. For such papers, grade boundaries are calculated that give as closely as possible the same percentages of candidates within each grade as those established judgmentally on the most closely associated examination paper.

Reviewing statistics on outcomes

It could be argued that in a criterion-related system dependent on professional judgment, the senior examiners should be able to set grade boundaries purely by considering the questions on the examination paper and what each question requires from the candidates by way of a response. However, in reality it is very difficult to make these judgments with any precision without reference to how candidates have actually responded. Cresswell (2000) concluded that awarders were typically correct in identifying when papers are easier or harder from one exam session to the next but not at estimating how much easier or harder they are.

> ... there are good theoretical and empirical reasons to believe that 'maintaining standards under the weak criterion-referenced definition' is too complex for even experienced awarders. Moreover there is empirical evidence that awarders' evaluative judgments are swayed by factors that should not influence them, such as consistency of marks across a script. These constitute good reasons to doubt that grading judgments made by awarders will be good enough, on their own, to maintain examination judgments.
>
> (Baird, Cresswell and Newton 2000: 213–229)

As the purpose of a grade award is to maintain a consistent meaning/standard for the grade, it is worth reflecting on one of Cresswell's definitions of comparability.

> Two examinations might be defined as having comparable standards if two groups of candidates with the same distribution of ability and prior achievement who attended similar schools with identical entry policies, are taught by equally competent teachers and are equally motivated receive grades which are identically distributed after studying the respective syllabuses and taking the examinations.
>
> (Cresswell 1996: 57–84)

To support the grade award, senior examiners are provided with the "statistically recommended boundaries" (SRBs). These are defined as the grade boundaries that would give exactly the same cumulative percentage as last year up to that grade. Cumulative percentage means the total getting that grade or

higher and is used so that if the most able candidates are performing more strongly (so the proportion getting grade 7 increases) we can account for the fact that we would expect fewer grade 6 as a result.

Examiners involved in the grade award process are also provided with information on the mean (average) mark obtained by candidates and histograms of the actual distribution of marks. While this information is taken into account by the SRBs it is often useful to be able to consider this greater level of detail.

Reflecting on the two statements above, it is important to remember that it will be a different group of candidates taking examinations each session. Therefore, it is likely that the assumptions of the second quote will not be completely met. This is particularly true if the two cohorts are very dissimilar to each other. Equally, when dealing with large numbers of students from similar educational experiences (such as an IB authorized school), it is more likely that significant differences in outcomes are due to them answering a different set of questions rather than variation in their performance.

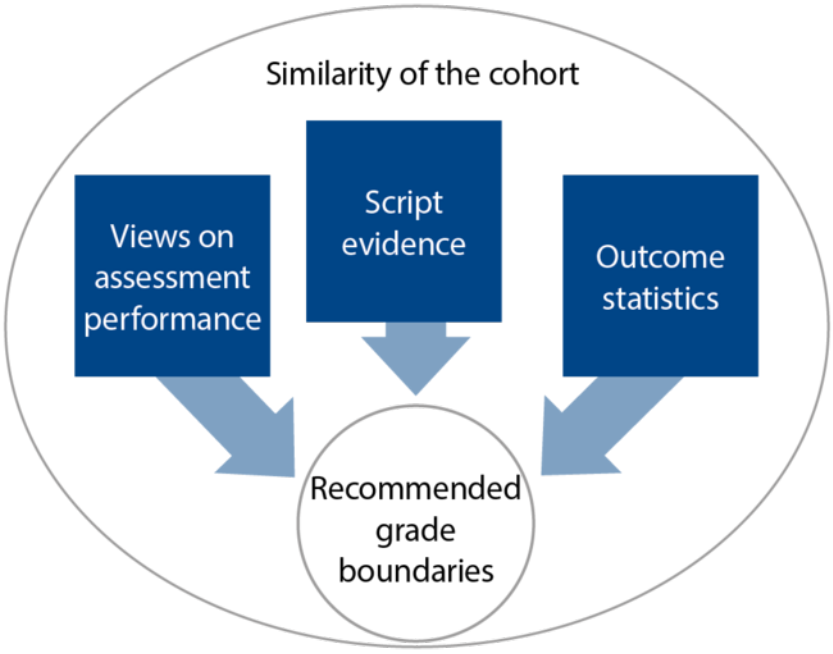
Cohort size must be taken into account when considering the significance of the SRBs; and the IB does not have formal rules around this as other factors are also important. In general terms, if there are only tens of candidates then we might expect considerable variation in the overall outcome, but if there are thousands of candidates, then this is relatively unlikely.

Balancing the evidence

No one type of evidence is more important than any other, and the task of the CE and their team is to balance them all together in making their recommendation.

Figure 57

Evidence that supports the selection of grade boundaries



Often, the different evidence suggests the same outcomes, for example, the SRBs lie in the zone of uncertainty; but sometimes there is a contradiction between them. In such cases, it is essential to explain the discrepancy when justifying the final decision.

At this final stage of grade award, it is possible to model what effect different decisions would have on the overall outcomes for the cohort. Recalling the maxim that it is the overall result that is important, not individual components, CEs might compare how the cohort's mean grade compares with teachers' predictions (referring to previous years to understand how reliable such comparisons have been). Another

approach that might be used is the exclusion of new schools from the overall results to see if they are performing very differently to more established institutions.

Finally, the CE needs to submit his or her recommendations to the IB's chief assessment officer together with the justification for these recommendations.

Grade descriptors

Grade descriptors are a compilation of the characteristics of performance at each grade. They play a critical role in setting grade boundaries as they describe what examiners need to look for when reviewing candidate work. This means they need to be written at the subject level to ensure they are applied consistently between examination sessions.

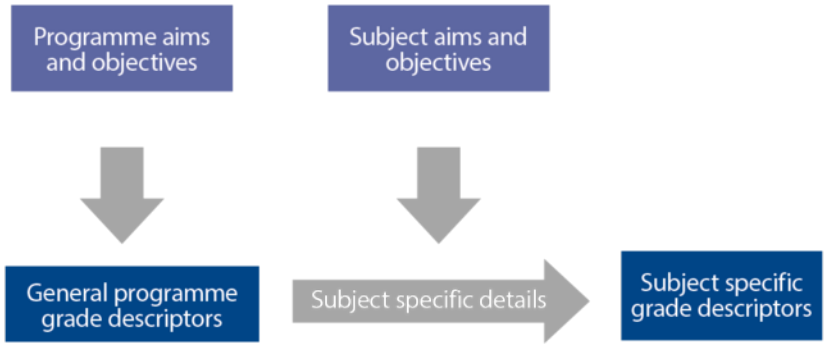
The importance of this role means that grade descriptors are an essential part of the validity of the assessment of our courses. They need to reflect the aims and objectives of the subject to ensure we are assessing and rewarding what is intended.

They also play a critical role in supporting inter-subject comparability by acting as a common benchmark for all subjects. Therefore, there is a need for general grade descriptors so that, for example, a grade 4 has a common meaning across all subjects.

These principles mean there should be a clear link between the goals of the IB programme and the general grade descriptors as well as clarity in moving to subject specific grade descriptors. The diagram below is a representation of this.

Figure 58

Relationship between general and subject grade descriptors



Fixed grade boundaries

- Where tasks are the same every session, the expectation is that grade boundaries will remain the same.
- Where there is evidence that standards are incorrect or have shifted over time the IB will review and change these "fixed" boundaries.

For many internally assessed tasks, and for some externally marked components, the task that the candidates are asked to complete is essentially the same for every session. Examples might include preparation of an art portfolio or a personal project. In such cases, it is reasonable to assume that, as the IB maintains the same marking standard every year (through standardization with previous years' work), the grade boundaries will also remain consistent.

During the grade award process, the PE and CE will be asked to consider whether there is any evidence that the existing grade boundaries are not appropriate, and in most cases we would expect them to conclude that the existing boundaries should be carried forward.

Despite this, it is important to be clear that these boundaries are considered every year, and that they can be varied if there is evidence that they are no longer effective in maintaining the comparability of grades between sessions. An example of how this could occur is when there is a step change in candidates' behaviour, such as new developments in technology for completing a task, changes in the work submitted which increases alignment with the markscheme without increasing the overall quality of candidate work, or a revision to the approach IB takes in marking or moderating work which results in a change of marking standard.

Aggregation

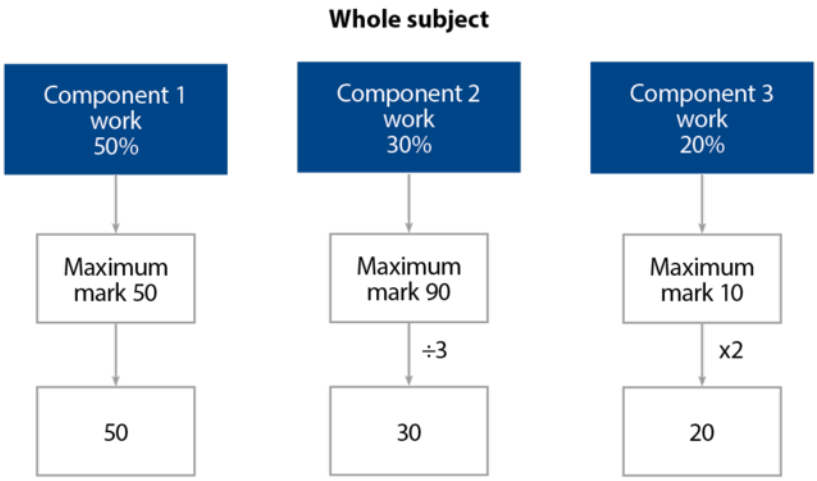
- Aggregation is the process of combining components to generate an overall result.
- In order for each component to contribute in the correct proportion to the final mark (weighting) it may be necessary to scale the component marks.
- The IB uses a "compensation model" where candidates can offset poor performance in one question or component with high attainment in another.
- A candidate's final subject grade is determined from the aggregation of component marks, and not from component grades.

Aggregation is the process of combining marks (and boundaries) from the different components to form a final mark or overall grade boundary. To achieve this, overall component marks (or boundaries) may need to be scaled.

Scaling is carried out to preserve the desired weighting for each component in terms of its contribution to the overall assessment for the course. It means multiplying or dividing component marks so that they contribute correctly to the overall total for the subject. The same applies to the grade boundaries set for the component, which would have to be determined initially out of the maximum marks for the components.

Figure 59

Example of scaling in a subject with three components



The concept of weighting is to reflect the relative importance that the IB places on the elements being assessed in contributing to the final outcome. For example, if a component primarily tests interpretation of data or sources, and has a weighting of 30%, then this implies that, compared with the other objectives of the course, interpretation represents about 30% of what is important. Often several components will test similar objectives and so this calculation is less meaningful.

A secondary, but important, aspect of weighting is that it allows us to set the total number of marks in an assessment that is appropriate to the tasks and the marking criteria rather than trying to force it artificially into an overall total.

It is a very important point that we do not require an individual candidate's marks to match the component weighting. We recognize that different candidates have different strengths, and this is why we use a variety of assessment instruments. It is important that we review the actual weightings of the entire cohort's results against those set out in the design of the subject. If one component's contribution is much higher in a session than intended, then this may indicate that the paper was particularly easy and we should look at other evidence in grade award. Alternatively, if one component's contribution is much higher than intended on a regular basis, then we need to review the design as the assessment is not working the way it was intended.

The approach outlined may not reflect the more sophisticated methods of weighting, combining (aggregating) and scaling described by, for example, Wood (1991, Chapter 10), but is based on sound criterion-related principles and supports the IB's principles of assessment.

This approach to aggregating the final mark means that a candidate can offset poor performance in one component with high performance in another, as it is only the total mark which impacts on the final grade. This concept is referred to as a "compensation model" to contrast it with a "mastery model" where a candidate would need to show the required level in all components to be awarded that grade.

It is worth stressing that a candidate's final subject grade is determined from the aggregation of component marks, and not from component grades. Because each component grade represents a range of marks, it is quite possible for two candidates with the same component grades to be awarded different subject grades.

Quality checks on grade awards and distribution reports

- Determining where grade boundaries should be set is governed by processes to support the IB in reaching a balanced decision.
- The CE and senior examiner teams make a recommendation on the grade boundaries.
- The IB subject manager is responsible for checking the grade award has followed IB's principles and processes.
- The recommendation and justification (including key data) is recorded in a 24-hour distribution subject report which is scrutinized by the IB Assessment leadership team.

The grade award process has been developed to support judgments made by examiners, by minimizing any unintended bias or examiners making decisions based on atypical work. The role of the IB's subject manager in supervising the grade award is to ensure that this process has been followed correctly and to support the senior examiners in identifying the challenges that will arise.

The initial recommendations for where to place the grade boundaries are made by the CE and his or her senior team. These recommendations must be supported by a robust argument setting out why these boundaries are the most appropriate based on the available (and perhaps contradictory) evidence.

The IB subject manager will act as a critical friend during this process, to ensure that the recommendations made by the CE are balanced and justified.

This recommendation and its justification is then presented in the 24-hour distribution subject report. This report also contains the underlying data such as changes in the cohort taking the course and grade outcomes from the proposed boundaries compared with previous years. The 24-hour report is then reviewed by senior members of the Assessment Division (usually the Programme Head and Head of Assessment Principles and Practices) to determine whether the arguments are sufficiently robust.

Where the IB Assessment leadership team is not happy with the recommendations made, they will discuss their concerns with the CE and ask them to place greater emphasis on one of the aspects of evidence or to provide more analysis to support their recommendations.

The IB's Chief Assessment Officer has the final authority on where grade boundaries are placed, based on the recommendations from the CEs.

Awarding a programme certificate

The IB programme certificates (IB diploma, CP certificate, MYP certificate) are not awarded through a grade award process. They are determined during preparation for publication of results and the process is described in the "Preparation for release of results".

Teacher observers

The IB is committed to increasing the transparency of its assessment processes and increasing general understanding of how grades are awarded. As part of this commitment teacher observers are invited to attend grade award meetings (precise details will depend if the meeting is face to face or being held virtually). There is an expectation that teacher observers will report back to colleagues on their experience and provide a report for the wider IBEN community. For more details please contact assessment.info@ibo.org.

Principles of grade award

The underlying principles of the IB grade awarding are:

1. The 3/4, 6/7, and 2/3 grade boundaries are determined (in that order) using all the available evidence (judgmental and statistical). Where there is no candidate work to establish these grade boundaries, the evidence that is available will be used to establish whichever boundary is most appropriate.
2. The other grade boundaries are then determined arithmetically according to the appropriate procedure.
3. Grade boundary decisions are made based on a triangulation of evidence from examiner judgment, statistical evidence and cohort information. All of these must be balanced equally and a compromise established.
4. If the cohort for an assessment is broadly similar to previous years, then we would expect the outcomes to be broadly similar to previous years. But:
    a. cohorts often do vary between years, particularly in small entry subjects
    b. where the outcomes do vary we would expect strong evidence to understand why.
5. If the tasks of an assessment are broadly similar to previous years then we would expect the grade boundaries to be similar to previous years.
    a. All grade boundaries can change between years, even for IA tasks.
    b. While we make every effort to ensure consistency in difficulty of assessments between years, we recognize that the demand of particular papers will vary.
6. It is the overall course grade boundaries that are the priority. It is this outcome that is significant to the candidate and is used by stakeholders to make decisions.
7. Component grade boundaries are a key step to arriving at a robust overall course result but small effects at component level can combine to have a large effect on the whole course outcomes.

"At risk" based quality checks

- The most important outcome of assessment is the grade the candidate receives so the final checks focus on this.
- The purpose of these final checks is to look for anything that appears unusual—if we had concerns with the marking this would have been addressed before this stage.
- If we identify any patterns of changes to marks in this process, we will investigate and possibly re-mark all of an examiner's work.
- These checks are done by our most consistent examiners, and their mark (if different) is more appropriate than the earlier score.

The quality model for marking works by periodically checking the standard of our examiners (through seed scripts) so we can be confident that they are marking their other scripts correctly. An alternative way of reviewing quality is to look for unexpected results and check they are correct—we call this "at risk" reviews.

We currently use two criteria to identify unusual results:

- Where an individual candidate has achieved a grade much lower than that predicted by the teacher.
- Where the overall results for a school are very different to last year. We particularly focus on cases where it is one component that looks very different to previous years.

In neither case does this mean that the marks awarded are incorrect, but it does indicate areas where we may want to do an additional quality check. In general, we prioritize those cases where there is evidence that candidates have done worse than expected rather than better.

We only use those examiners who have shown they are the most consistent (through the seeding quality model) so we have confidence that the mark they award is correct. For this reason, if they suggest a different mark, we use their judgment rather than the earlier examiner. However, if we see a large difference between two consistent examiners, we will investigate further and may seek a third opinion to understand why there is a difference.

In e-marking all these marks are recorded by the computer, but in the case of paper scripts schools may occasionally see two or more sets of marks as a result of this review process. The IB will always explain which marks relate to the most senior and reliable examiner.

The outcome of the "at risk" marking is twofold:

1. To check that candidates who have received an unusual result are receiving a fair outcome.
2. To look for any evidence of a systematic problem with the marking, for example, a particular question which several examiners found difficult to mark, or particular examiners who were not consistent in their marking. In such cases we would re-mark the affected candidates' work.

The final award committee

- The final awards committee (FAC) is the strategic-decision making body for each examination session.
- It consists of senior IB staff from across the organization and CEs from several different subject areas.
- As well as giving final approval for the issue of results, it also considers cases of academic misconduct, special consideration and any other pertinent issue.
- As well as its voting members, schools and/or stakeholders can be invited to observe its proceedings.

The last stage in the approval of candidates' results is the Final Awards Committee (FAC). It is this body that reviews the IB Chief Assessment Officer's recommendation that the session has met the standards of the IB and results should be awarded on behalf of the Board of Governors. It also sets policy and precedents relating to the awarding of IB qualifications.

Its precise remit and composition varies slightly between the different programmes, but in general it consists of equal numbers of voting members drawn from:

- senior IB staff covering assessment, academic and school services divisions
- CEs from a range of different subject areas.

Unlike the checks carried out by senior assessment staff on the 24-hour distribution reports, the FAC acts as an oversight board and reviews the macro-level outcomes such as overall programme completion rates and any issues brought to its attention by staff in the IB Assessment division.

In a similar way, the FAC reviews issues of academic misconduct and maladministration. Recommendations are made by a sub-committee of the FAC and discussed, establishing precedents in new cases.

Requests for special considerations are treated in a similar way with staff in the IB Assessment division making recommendations to the FAC, which are discussed before reaching a final decision.

The final role of the FAC is to reflect on the performance of the examination session and make recommendations to the IB for subsequent sessions.

Conflict of interest

The FAC is an important decision making body and so it is critical that it is seen to be both transparent and independent. The policy is that any member who could be perceived as having a conflict of interest on a particular agenda item will leave the room for that discussion, and this is reiterated at the start of every meeting.

Observers

The IB encourages observers to attend the FAC meetings in order to:

- make the procedures of the final award committee more transparent
- provide an opportunity for feedback on the procedures of the final award committee
- acknowledge the partnership between IB constituents
- provide an opportunity for suggestions for change and improvements to the procedures of the final award committee through the submission of a written report.

In order to support the IB in meeting these objectives, we ask observers to submit a report on their attendance to the Chief Assessment Officer within two weeks of the meeting. This report should include

general observations on current procedure and process, and suggested changes and improvements, where appropriate. No comments should be made on individual cases that were considered by the committee.

Due to the sensitive nature of the issues discussed at FAC, observers are also bound by the appropriate restriction on confidentiality and personal issues of conflict of interest. Observers have no voting rights on the decisions of the committee.

For more details on being an observer at an FAC meeting please contact assessment.info@ibo.org.

Preparation for release of results

- Once all scripts have been marked and grade boundaries determined, there are several processes that need to be completed to ensure that candidates' results are ready to be published.
- Where there is no evidence of candidate work the IB may estimate a mark using the missing mark procedure.
- Candidates' results need to be combined in order to determine whether they have achieved the DP, CP or MYP certificate.
- Any candidates granted special considerations need to be reviewed.
- The publication of results is primarily about implementing a robust change control system so that any alteration to grades are picked up and the relevant stakeholders, primarily schools and universities are informed.

Between setting grade boundaries and issuing results to candidates there are a number of processes and procedures that need to be completed. Some of these only apply to certain candidates, such as carrying forward anticipated subjects or implementing special considerations, others affect all candidates.

Missing marks

- Where the IB is not able to access the candidate's work due to no fault of the school or the candidate, we will estimate a mark to minimize the potential disadvantage to the candidate.
- Such an estimation must be based on evidence, in particular, if we have very little candidate work to base an estimate on, an alternative solution to "missing mark" must be found.
- The missing mark process is based on average candidate performance, therefore roughly equal numbers of candidates will be advantaged by its estimation as are disadvantaged. The fairest result is always to have actual candidate work to mark.

There are sometimes cases where a candidate's work is not available to the IB to mark for reasons beyond the control of the school and the candidate. Examples include when examination papers are lost in the post or where candidates have a sudden illness on the day of the examination.

The Missing Mark Procedure (MMP) is a mechanism that we can use in such circumstances to estimate a mark for the candidate so that they are not unfairly disadvantaged.

It is appropriate to invoke this procedure in circumstances where the reason for the lack of the candidate's work is due to the actions of the IB or third parties (not including the school) where it would not be reasonable for the candidate to be asked to complete the assessment on another occasion.

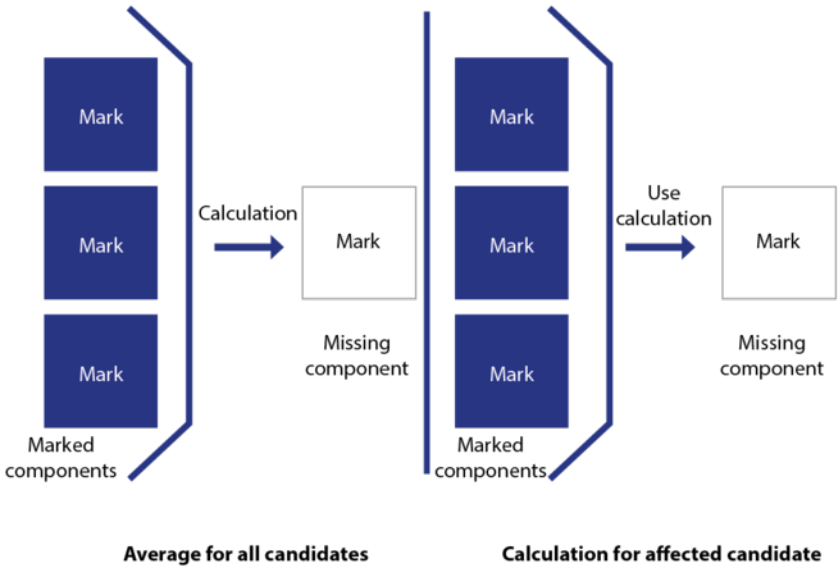
MMP is not intended as a substitute for access arrangements or special considerations (in all but exceptional cases).

In all cases, the MMP must be based on evidence of the candidate's achievement, so if we do not have much information on the candidate's performance it will be very unlikely we can give a fair estimated mark.

The MMP is based on the average attainment of all candidates in that component compared with their performance in the other components. It is important to recognize that by taking this average we would expect as many candidates to perform worse by this estimation as would perform better. It is therefore always a fairer result if we can use actual candidate's work to mark.

Figure 60

How the missing mark procedure works



If there is only one component in a particular subject's assessment model then the MMP cannot be used—there is no evidence to use to estimate a mark. This is the case in the MYP and so we employ a different approach—the MYP missing grade procedure.

Programme outcomes

- The IB awards diplomas and certificates on the basis of meeting criteria, not on individual judgments.
- Each programme has its own criteria for the award of the certificate

Once we have all subject results from a candidate, we can calculate whether they have qualified for the DP certificate, CP certificate and MYP certificate. This is based on the individual meeting criteria which are detailed in the relevant sections.

How the diploma outcome is calculated

The overall diploma points are calculated by adding together the grades (1 up to 7) achieved from each of the six subjects and then including between zero and three points from the core. This means that the highest score that a candidate can achieve is 45 points*.

This approach means that SL and HL subjects are valued equally in determining the candidate's final points.

*The maximum points of 45 is obtained from 6 (subjects) times 7 (top grade) plus 3 points from the core.

Core points matrix

Unlike the other subjects, theory of knowledge (TOK) and the extended essay (EE) are graded from A to E. The third element of the core, CAS, does not receive a grade as it would not be meaningful to evaluate performance in this area.

The core is worth between zero and three points towards the overall diploma points. The candidate can also fail to achieve the diploma certificate if they obtain a grade E in either TOK or EE or if they do not complete CAS. The number of points is calculated using the table below.

Figure 65

Core points matrix

	Theory of knowledge (TOK)					
Extended essay	Grade awarded	A	B	C	D	E
	A	3	3	2	2	Failing condition
	B	3	2	2	1	
	C	2	2	1	0	
	D	2	1	0	0	
	E	Failing condition				

Failure conditions

A candidate can only receive the overall diploma certificate if none of the following nine conditions below applies.

- CAS requirements have not been met.
- Candidate's total points are fewer than 24.
- An N (no grade awarded) has been given for theory of knowledge, extended essay or for a contributing subject.
- A grade E has been awarded for one or both of theory of knowledge and the extended essay.
- There is a grade 1 awarded in a subject/level.
- Grade 2 has been awarded three or more times (HL or SL).
- Grade 3 or below has been awarded four or more times (HL or SL).
- Candidate has gained fewer than 12 points on HL subjects (for candidates who register for four HL subjects, the three highest grades count).
- Candidate has gained fewer than 9 points on SL subjects (candidates who register for two SL subjects must gain at least 5 points at SL).

Bilingual diplomas

As an alternative to the standard diploma certificate, a "bilingual diploma certificate" can be awarded to a candidate who:

- completes two languages selected from group 1 with the award of a grade 3 or higher in both
- completes one of the subjects from group 3 or group 4 in a language that is not the same as the candidate's nominated group 1 language. The candidate must attain a grade 3 or higher in both the group 1 language and the subject from group 3 or 4.

Pilot subjects and interdisciplinary subjects can contribute to the award of a bilingual diploma certificate, provided the above conditions are met.

The following cannot contribute to the award of a bilingual diploma certificate:

- an extended essay
- a school-based syllabus
- a subject taken by a candidate in addition to the six subjects for the diploma certificate ("additional subjects").

How the CP outcome is calculated

There is no points score associated with the CP certificate.

The CP certificate will be awarded to a candidate provided all of the following requirements have been met.

- The school has confirmed that the candidate has completed the specified career-related study.
- The candidate has been awarded a grade 3 or more in at least two of the DP courses.
- The candidate has been awarded at least a D grade for the reflective project.
- The school has confirmed that all personal and professional skills, service learning and language development requirements have been met.
- The candidate has not received a penalty for academic misconduct from the final award committee.
- The career-related diplomas and reflective project grades are confirmed by the same final award committee as the DP.

Bilingual CP certificates

In addition to the usual certificate, a "bilingual certificate" can be awarded to a candidate who:

- completes two DP language courses selected from studies in language and literature with the award of a grade 3 or higher in both
- completes a DP language course from studies in language and literature and also completes a DP course from individuals and societies or sciences in a response language that is not the same as that taken from studies in language and literature. The candidate must attain a grade 3 or higher in both courses.

How the MYP outcome is calculated

At the end of their year 5 MYP studies, candidates can be entered for the IB external assessment. The outcomes of these assessments will be recorded in an MYP Course Results document. In addition, these candidates can choose to take assessments which can lead to the award of the MYP certificate.

The school can also issue an MYP Record of Participation. This is for MYP students who study the programme for at least two years and complete the requirements in year 3 or year 4. These students are not registered with the IB for any form of assessment. The Record of Participation is a school-based document, not verified by the IB.

In order to achieve the IB MYP certificate, the student must have participated in the final year of the programme, with a recommended period of participation of two years, and:

- complete either an on-screen assessment or ePortfolio in six subjects consisting of: language and literature, language acquisition (or a second language and literature), individuals and societies, mathematics, sciences and one subject from arts, physical and health education or design
- achieve at least a grade 3 in each of the six subjects above
- complete the on-screen examination in interdisciplinary assessment and achieve at least a grade 3
- complete the personal project with at least a grade 3
- obtain a total of 28 points overall
- meet the school's expectations for community service.

The MYP bilingual certificate additionally requires successful results from on-screen examinations for one of the following:

- a second language and literature course (instead of a course in language acquisition)
- one (or more) science, individual and societies, or interdisciplinary examination in a language other than the student's chosen language and literature course.

Fairness for all—meeting candidates' needs

In order for our assessments to be valid they must not discriminate against candidates with particular needs. The IB will consider requests for modified papers and inclusive arrangements as set out in the programme's *Assessment procedures*.

- The best way to ensure fairness is through designing assessment to be accessible for everyone, thus removing the need for any modification. This is encapsulated in the concept of Universal Design of Assessment, part of IB's commitment to Universal Design for Learning (UDL).
- Some candidates' needs are known about in advance and these are dealt with through inclusive access arrangements which may include modified papers.
- Other circumstances arise at short notice or cannot be managed through inclusive arrangements. In these cases, we treat them through our special consideration processes.
- Ultimately, the purpose of all these arrangements is to create fairness for all our candidates, and so in reaching any decision the IB must consider what is fair for the entire cohort and not just the one individual candidate. The aim is to have an even playing field for every candidate.

The IB believes that all candidates should be allowed to demonstrate their ability under assessment conditions that are as fair as possible. We recognize that standard assessment conditions may put candidates with learning support requirements at a disadvantage by preventing them from demonstrating their level of attainment. Similarly, we acknowledge that sometimes events or circumstances beyond the control of the candidates will affect their performance and should be taken into account.

The best way to ensure fairness with an assessment is for everyone to take the same assessment in the same way. Many of the modifications made to support specific requirements would help all candidates in understanding and engaging with the questions. The ideal situation is for all assessments to be developed with an understanding of the range of requirements that candidates may have rather than to treat some candidates differently. This is the concept of Universal Design of Assessment. The IB recognizes that this total inclusivity approach is sometimes not achievable and so we also have a process for requesting specific inclusive arrangements.

Inclusive access arrangements are designed to meet candidates' individual needs, such as:

- learning disabilities
- language difficulties
- specific learning difficulties
- communication and speech difficulties
- autism spectrum disorders
- social, emotional and behaviour challenges
- multiple disabilities and/or physical, sensory, medical or mental health issues.

Any reasonable adjustments for a particular candidate pertaining to his or her unique needs will be considered. For further details, please refer to *Assessment procedures* and the IB publication *Candidates with assessment access requirements.*

Adverse circumstances are those that are beyond the control of the candidate and which might have a negative impact on their performance. Such cases are considered by the Final Awards Committee and if accepted candidates close to a grade boundary will receive the higher grade.

The accepted IB principle of fairness to all candidates means that, when considering any inclusive arrangement or adverse circumstance, we should not create a situation that is unfair for other candidates taking the assessments. The goal is for a level playing field for all candidates.

Principles for inclusive access arrangements

The principles for inclusive access arrangements are set out in the IB document *Candidates with assessment access requirements*. The text below is taken from the DP version, but similar principles apply to other programmes.

1. The IB must ensure that a grade awarded to a candidate in any subject is not a misleading description of that candidate's level of attainment, so the same standards of assessment are applied to all candidates, regardless of whether or not they have learning support requirements.
2. Inclusive access arrangements, including reasonable adjustments, are pre-examination measures for a candidate to access the assessment. They cannot be requested retrospectively either for oral or written examinations.
3. The arrangements requested for a candidate must not give that candidate an advantage in any assessment component.
4. The inclusive access arrangements described in this document are intended for candidates with the aptitude to meet all assessment requirements leading to the award of the diploma or course results.
5. When inclusive access arrangements are necessary for a candidate during the course of his or her study of the Diploma Programme or practice examinations, the school may provide the arrangements. If the arrangements are required for assessment, this document lists the arrangements that do not require prior authorization from the IB. For all other arrangements, prior authorization from the IB Global Centre, Cardiff is mandatory. Similarly, if a Diploma Programme candidate has difficulties meeting the requirements for creativity, activity, service (CAS), IB Answers must be consulted.
6. Schools are advised to plan inclusive access arrangements for their candidates based on the IB criteria as stated in this policy and teachers' observations of the candidate in the classroom during class work and tests.
7. The inclusive access arrangements requested for a candidate must be his or her usual way of working during his or her course of study. Only in very exceptional and unusual cases, will the IB authorize a request for inclusive access arrangements that are not the usual way of working and that have been put in place to support the candidate only in the last six months of study or thereafter, just prior to the examinations.
8. The IB aims to authorize inclusive access arrangements that are compatible with those normally available to the candidate concerned. However, authorization will only be given for arrangements that are consistent with the policy and practice of the IB. It should not be assumed that the IB will necessarily agree to the arrangements requested by a school. Coordinators are required to provide information on the candidate's usual method of working in the classroom.
9. The IB is committed to an educational philosophy based on international-mindedness. Therefore, the inclusive access arrangements policy of the IB may not reflect the standard practice of any one country. To achieve equity among candidates with assessment access requirements, the policy represents the result of a consideration of accepted practice in different countries.
10. The IB will ensure that, wherever possible, arrangements for candidates with a similar type of access requirement are the same. Due to the cultural differences that occur in the recognition of learning support requirements and the nature of access arrangements granted in schools, there may be some compromise that may be necessary to help ensure comparability between candidates in different countries.
11. Each request for inclusive access arrangements will be judged on its own merit. Previous authorization of arrangements, either by the IB or another awarding body, will not influence the decision on whether to authorize the arrangements that have been requested by the coordinator.
12. The IB treats all information about a candidate as confidential. If required, information will only be shared with appropriate IB personnel and members of the final award committee, who will be instructed to treat such information as confidential.

13. If a school does not meet the conditions specified by the IB when administering inclusive access arrangements or makes arrangements without authorization, the candidate may not be awarded a grade in the subject and level concerned.
14. If it can be demonstrated that a candidate's lack of proficiency in his or her response language(s) arises from an identified learning support requirement, inclusive access arrangements may be authorized. (For subjects in groups 3 to 6, all candidates are allowed to use a bilingual/translation dictionary in the written examinations.)
15. A school must not inform an examiner of a candidate's challenges (such as autism, writing difficulties and so on) or adverse circumstance.
16. In the case of internally assessed work, teachers must not make any adjustments when marking a candidate's work.
17. The list of inclusive access arrangements available is revised regularly. The IB will consider alternative arrangements proposed by a coordinator, provided those arrangements could be made available to all candidates with similar requirements.

According to the document *General regulations: Diploma Programme*, a Diploma Programme candidate may participate in three examination sessions to be awarded the diploma. At the discretion of the IB, a candidate with learning support requirements may be allowed additional sessions.

1. If the nature of a candidate's challenge and/or the authorized inclusive access arrangement might disturb other candidates during an examination, the candidate must take the examination in a separate room and be supervised according to the regulations governing the conduct of Diploma Programme examinations.
2. Written examinations must be invigilated according to the regulations governing the conduct of Diploma Programme examinations. The person invigilating the candidate's examination must not be a relative of the candidate, or any other person with whom there may be an apparent or perceived conflict of interest.
3. Any issues that arise from the nature of the inclusive access arrangements, or any unforeseen difficulties encountered by the candidate during the examinations, should be reported to IB Answers as soon as possible.

Exemptions from assessment

Exemptions are not normally granted for any assessment component. However, if an assessment component or part demands a physiological function that a candidate is not able to perform, an exemption may be authorized. Before submitting a request for an exemption from a component, careful consideration should be given to whether all reasonable adjustments have been considered. Authorization for an exemption will only be given when there are substantial grounds for an exemption. A candidate's physical inability to perform the functions required by the component must be clearly and fully documented.

For full details on the principles and processes around exemptions from assessment please refer to the *Candidates with assessment access requirements* document for the appropriate programme.

Opportunities for inclusive arrangements with on-screen assessment

The use of on-screen assessment allows the candidate to take far more control over how they wish the assessment to be presented. Computers are able to provide a large variety of fonts, text sizes and colours to meet an individual's needs and this type of adjustment can be routinely available to every candidate.

In many current cases, the inclusion arrangement requested is to allow the use of a computer and this need is clearly met by eAssessment so long as the inclusion software required is compatible with the on-screen tool. The IB is very aware of this requirement and is working to ensure that any on-screen examinations meet the industry standards for such inclusion software.

Adverse circumstances

Adverse or unforeseen circumstances are those that are beyond the control of the candidate and which might have a negative impact on his or her performance. This includes temporary illness or injury, severe stress, exceptionally difficult family circumstances, bereavement, or events that may threaten the health or safety of a candidate. Adverse circumstances may also include an event that affects the whole school community, such as civil unrest or a natural disaster.

Adverse circumstances do not include shortcomings on the part of the school. It is a school's responsibility to ensure that all candidates comply with programme and assessment requirements, including issues with teaching staff.

Full details of what is included and excluded within the category of adverse circumstances can be found in the appropriate programme's *Assessment procedures* and *General regulations*.

In such cases, the evidence supplied by the school will be considered by the final award committee, to determine if the candidate affected should be eligible for special consideration. If a candidate's circumstances are deemed "adverse" and therefore qualify for consideration, an adjustment may be made to the candidate's total mark in the affected subjects or programme core requirements. If the candidate is within one or two scaled marks of the next higher grade boundary, the candidate's grade in the affected subjects will be raised.

Universal design of assessment

The preceding text has discussed how the IB manages situations where candidates need modifications or specific assistance to be able to fairly take our assessments. The best solution, however, is to have assessments which do not have these barriers to participation in the first place. The concept of Universal Design of Assessment is to consider access, inclusion, equality, cultural sensitivities, stereotypes and bias from the starting design of the assessment. This includes the creation of examination tasks and questions, but also goes a step back into the design of the overall assessment model which sets the framework of how comparable assessments are created for every session. By creating more inclusive, and indeed less construct irrelevant, assessments at the start we can minimize the challenges faced in meeting the needs of individual candidates.

Universal Design of Assessment is an aspect of the overall Universal Design for Learning (UDL). UDL focuses on creating accessible learning environments for all learners, including candidates with disabilities, candidates from culturally and linguistically diverse backgrounds and candidates who are gifted and talented. The principles of UDL are relevant across the education including in curriculum design, school management and teaching. For more details on UDL in the IB, refer to Rao K, Currie-Rubin, R and Logli C. 2016. *UDL and Inclusive Practices in IB Schools Worldwide*.

Publication of results

From the perspective of the IB, the publication of results is primarily about implementing a very strict change control protocol.

During the examination session, information is constantly updated so we have a clear picture of what is going on. Once candidates and schools have been informed of their outcomes we need to be very clear that they do not change without everyone being properly informed. For the DP and CP this particularly applies to universities, which receive transcripts of candidate's outcomes.

Authorized changes to results after they have been issued may be due to EURs, confirmation of pending results from schools or the resolution of academic honesty cases.

Enquiries upon results (EUR), appeals and general feedback

- The purpose of the enquiries upon results (EUR) service is to allow schools to highlight to the IB where they believe a mistake has been made in the marking process.
- Exactly the same standards must be applied during EURs as during the main examination session, and the IB uses its seeding quality model to ensure this happens.
- The external IB assessments are intended as summative rather than formative assessments and so we require our examiners to only write comments when it supports their marking.
- The IB also has a formal appeals process if schools or candidates do not believe that the correct processes have been followed.

For the legal and procedural description of the EUR services offered, please refer to the relevant programme's *Assessment procedures*.

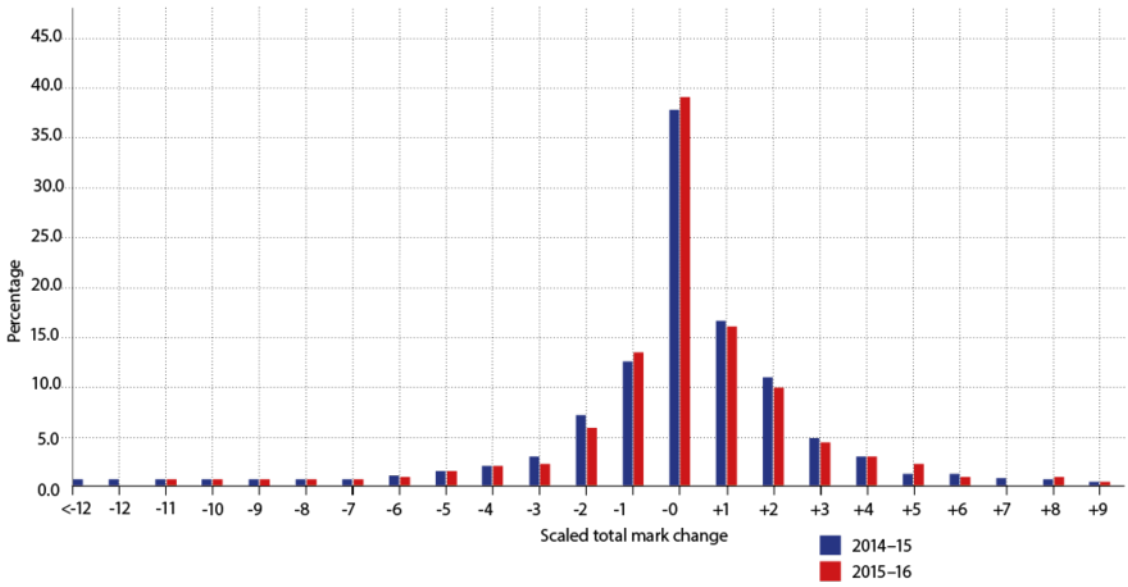
Purpose of EUR

The EUR process is intended as a final safeguard against errors in the marking system. It is an opportunity for schools to highlight to us when they believe a mistake has been made and for us to investigate and, where necessary, correct such errors.

As has been described in the marking quality model section, two different examiners could have a small disagreement on the number of marks to award a particular question without either of them having made a mistake. Therefore, if the EUR process results in only a very small change in the number of marks then it is likely that there has not been a mistake in the marking. This is the case for the majority of our EUR requests.

Figure 61

Mark changes from EURs in May 2015 and 2016



In the EUR process, we only use our most consistent and senior examiners and so based on the hierarchy of examiners we regard the result of this re-mark as the correct mark.

Given the nature of setting grade boundaries, there is always the possibility of candidates who are one mark below the boundary, or just over it. In such cases an EUR outcome could result in a change in grade despite the previous comment on there being no mistake in the marking. In such cases the IB would argue that the candidate is on the boundary between two grades and either is a fair representation of their performance.

To minimize the impact of these small, and appropriate, differences in examiners' marks we re-mark all external examinations taken by a candidate during an EUR. This mitigates the impact of a small change as, assuming that any changes are not systematic, we would expect them to cancel out over several papers.

Maintaining standards

The purpose of EURs is to provide candidates with the marks that they should have received during the marking period. It is very important to ensure that examiners are not inappropriately influenced by the fact that the candidate is unhappy with their initial mark, or is close to a grade boundary.

To support examiners in maintaining this standard the IB includes seed scripts within the EUR work, which alert markers electronically if they are moving away from the agreed standard.

To mitigate against an EUR examiner making a serious error, the IB reviews every proposed EUR mark change and, when there is a large and unexplained difference between the two examiners, we will ask for a third opinion on the script. In such cases, we will generally use the mark provided by the most senior examiner.

Identifying systematic issues

The IB monitors all of the changes to marks during the EUR process to check whether there are any patterns that could indicate an issue with the original marks, for example, an examiner who had passed our quality checks but was not consistently marking to the same standard.

When we see any evidence of such systematic errors we will investigate them to establish if there is an underlying cause. If we believe that there is such a problem, we will proactively identify every candidate who has been affected (whether or not they have requested an EUR) and ensure they have not been disadvantaged.

It is also common for schools to raise concerns with us about systematic issues. We take these seriously and will review any evidence available, but under most circumstances will not undertake a re-mark without the school requesting it through the EUR process.

Categories of EURs

After the issue of results, the coordinator may request a:

- category 1—re-mark of all a candidate's externally-assessed components for a subject
- category 1 report—report on the marking of a category 1 EUR
- category 2—copies of externally-assessed component material
- category 3—re-moderation of an internally-assessed component.

A fee is payable for each of the above categories (except when a grade is changed as a consequence of a category 1 re-mark).

A returned script may contain useful comments from the allocated examiner, however, this can't be guaranteed because examiners are not required to write comments when marking candidates' work. This is because the purpose of summative assessment is to produce an accurate reflection of a candidate's performance at the end of his or her study and the emphasis for examiners is to make sure the work is marked correctly, rather than to provide notes and recommendations on the work itself. Feedback on

candidate performance is an important part of formative assessment which is carried out by the teacher throughout the course.

As mentioned in the "Moderation" section regarding of externally assessed work, the mark awarded by an examiner as shown on a script may not necessarily be the final "moderated" mark awarded and it is the mark awarded to the candidate that is shown on IBIS that is correct.

Fairness for all (grades can go both up and down)

- Candidates' marks can go down as well as up as a result of a category 1 EUR, therefore the candidate's permission must be sought for submitting work for the EUR.
- It would not be reasonable of the IB to require a school to obtain the permission of all candidates in a class before asking for an EUR category 3 (re-moderation). Therefore, in this case (only), candidate grades are protected and cannot go down.
- Where a mark changes the result for another reason, then we will consider whether to protect the candidate's grade on a case by case basis, but where there is no fault on the part of the candidate or school the presumption would be to not allow the grade to be reduced.

The purpose of the EUR system is to ensure that candidates receive the correct mark for their work, and so the IB must take every care to make sure that the mark that results from any EUR is as accurate as it can be. The IB aims to do this by only asking the most senior and consistent examiners to undertake EUR marking and reviewing outcomes to make sure they have the appropriate justification.

The IB believes that the EUR mark represents the most accurate representation of the candidate's work, and so this is the mark and grade awarded. This means that a candidate's result can go down as well as up as a result of a category 1 EUR. Schools must therefore obtain the candidate's permission before submitting an category 1 EUR.

For a category 3 EUR (re-moderation), any change in marks could apply to the entire cohort. If we applied the principle of seeking candidates' permission, this would mean that a single candidate could stop their class' work being re-moderated by not giving their approval or simply not responding to the request. This would clearly be unreasonable and so, to prevent candidates' grades being reduced without their formal acceptance of this risk, grades can only increase, not decrease, as a result of a category 3 EUR.

If a candidate's mark changes for any other reason, we will consider whether a grade should be protected on a case by case basis. There is a strong presumption that, if there is no fault on the part of the candidate or school, then the grade should be protected as the candidate may have made decisions based on this outcome which they would not be able to change.

Feedback from marked assessments

The assessments that are considered in this section are summative assessments. While this does not preclude them also being used for formative purposes, the IB needs to be careful to ensure that their validity is not undermined by other purposes.

The IB expects examiners to mark accurately and appropriately, to the standard set by the PE. They are only asked to include comments that support their marking, not to provide feedback for candidates on how they could improve. As teachers are aware, offering instructive feedback is a separate skill to summative marking and should ideally be tailored by a holistic knowledge of the candidate. This is in direct contradiction to the IB's expectation that examiners will be completely free of any candidate bias in their marking.

The IB appreciates that simply being able to reflect on what was done well in an assessment can be of help to candidates and teachers in reflecting and improving their skills and understanding, and so the IB is happy to provide details of the marking outcomes, but with the understanding that they are not expected to provide formative comments.

Appeals

The EUR process should ensure any errors in marking or assessment processes have been resolved. If a school or candidate remains concerned that their assessments have not been handled properly they can lodge a formal appeal. Appeals are only possible once all other routes have been exhausted.

Appeals are possible against:

1. results—when a school has reason to believe that a candidate's result(s) are inaccurate after all appropriate EUR procedures have been completed
2. a decision upholding academic misconduct, but not against the severity of a penalty
3. a decision in respect of special consideration—following a decision not to give special consideration to a candidate as a consequence of alleged adverse circumstances
4. an administrative decision not covered by one or more of the foregoing circumstances, which affects the results of one or more candidates.

There are two stages to the appeals process. In stage 1, a senior member of IB staff who has not previously been involved with the case will review the evidence and the decision. If the complainant is still not satisfied, then stage 2 involves a review by a panel of three people including a member who is external to the IB.

Full details of the appeals process can be found in the *General regulations*. For more details email appeals@ibo.org or contact IB Answers.

The IB also has an ombudsman whose role is to act as a mediator to "work with individuals and groups in an organization to explore and assist them in determining options to help resolve conflicts, problematic issues or concerns". This alternative route is available for schools or candidates who are unhappy with the assessment process.

Setting next year's assessments

- Assessment is considered as a cycle—it is important we learn from each session to continually improve our assessments.
- Papers and markschemes are developed up to two years before a session, so different evidence will be taken into account at different points in the life cycle of the assessment.
- The evidence that needs to be taken into account covers operational experiences as well as candidate outcomes (results) and teacher feedback.
- It is an important principle to the IB that we use the full range of tools available to understand and correct any issues of subject comparability, not just adjustments to grade boundaries.
- Any change which could affect teaching pedagogy should have a lead-in time equal to the length of the course (that is, two years).

How should evidence be used?

Self-reflection is an important part of the IB philosophy and is as important in the assessment process as it is for the candidates we are assessing. As a result of each session, we will have gained a range of evidence around:

- how candidates interpreted examination questions and how closely this matched our intentions
- whether individual questions performed in the way we had intended—had we judged the demand and difficulty correctly?
- how reliably examiners were able to mark each question—were there any questions that were particularly difficult to maintain the correct standard?
- whether the different components performed in the way we planned—were the actual weightings in line with those intended?
- whether there were problems in preparing or marking examinations
- whether teachers felt the paper reflected their expectations—was this a problem with a specific question?
- whether teachers felt the paper was easier or harder than previous sessions—why?

We analyse and use this information in different ways depending on the nature of the issue. If it relates to issues with the questions being set, we would generally adapt the style of examinations as they were being written (up to two years before they are taken by candidates) unless they were sufficiently severe to require prepared examinations to be rewritten.

In contrast, where evidence suggests that assessment processes or marking standards need to be adjusted, this can take place for the next session.

Inter-subject comparability

One of the most challenging aspects of validity is comparability. As the figure below shows, it is often not clear when it is reasonable to make comparisons between two different things, and the complexity in comparing subjects is equally demanding—how do you compare performance in literature and mathematics?

Figure 62

Which of these can we reasonably compare? Do we use a relative or absolute approach?

The situation is made more complicated by the fact that different candidates are better at some subjects than others—there is no "average" candidate we can use to define how difficult an assessment was. There is a wide range of research around measuring inter-subject comparability. See, for example, Ofqual (2015) *Inter-Subject Comparability: A Review of the Technical Literature: ISC Working Paper 2*.

IB programmes generally require students to take a broad curriculum, choosing subjects from particular groups. This means that we need to focus our attention on making subjects within these groups equally demanding so that every route through the programme is equally challenging. Examples of this might be being concerned that biology and physics are comparable, or geography and history.

Comparisons between subject groups is also important so that candidates with different strengths are treated equally.

The most important consideration that the IB takes is to balance the different forms of evidence together; no one source dominates our decision-making.

In some subjects, particularly the languages, it is possible for our examiners to compare work between subjects to compare the difficulty. This usually happens as part of the standardization process.

After each session, the IB calculates a subject pairs analysis. This approach is well documented (Nuttall et al 1974), and works by looking at candidates who have done both subjects and comparing their grades in each. For each subject, we can then calculate an average difference with all the other subjects and establish a rank order of difficulty. This method is not without limitations (see Ofqual 2015 or Coe et al 2008) but used in conjunction with other evidence represents a meaningful source of information.

Where the subject pairs analysis or subject expert views suggest that there is a difference in standards between subjects, this would be included in the discussions at grade award. For example, if a subject is known to be "difficult" from subject pairs analysis, we would expect examiners to err on the generous side in any judgmental decisions. It is an important principle to the IB that we use the full range of tools available to understand and correct any issues of subject comparability, and not just make adjustments to grade boundaries.

Making changes

An important principle when making any change to assessment is that candidates and teachers should not be disadvantaged. This means that, where an alteration would affect either teaching practices or preparing for examinations, we will only make changes for candidates who are about to start their course.

Supporting curriculum reviews

The evidence from how the assessment performed is not only used to improve future assessments, but also to support the wider curriculum review process. A formal report on assessment is part of every curriculum review, but in general IB curriculum managers will attend the first grade award of a new course to observe first-hand the experiences of the examining team.

On-screen assessments

On-screen assessment should be no different to traditional assessment in how we learn lessons from previous sessions' assessments. The opportunities presented by on-screen to offer innovative and novel questions are a great strength, but also necessitates a careful evaluation of such items to ensure they performed as expected.

Feedback to schools

- The purpose of IB summative assessment is to measure a candidate's performance and all our processes should be designed to maximize the validity of this outcome.
- Part of this is transparency, so that candidates and teachers can see that marks and grades have been awarded consistently and without bias.
- This is especially important with IA where the final award is based on a moderation of the teacher marks.
- IB summative assessment, including the EUR process, is not intended to provide guidance to schools on how to improve candidate outcomes, except as a by-product of data required for a valid assessment.
- The IB does provide other services to schools to support their teaching pedagogy and professional development outside of the assessment process.

Examiner comments

> One cannot manage too many affairs: like pumpkins in the water, one pops up while you try to hold down the other.
>
> (Chinese proverb)

The purpose of IB summative assessment is to measure a candidate's performance and we require examiners to focus solely on marking candidates' work to the required standard. We only ask examiners to make comments when it helps them in doing their marking.

Writing formative feedback for either candidates or teachers requires the examiner to determine the correct mark for a piece of work and then try to explain how the work could have been improved. We reflect that the second task is at the heart of what good teaching means, and is not a trivial task. It requires time and thought which will draw the examiner away from their core task of marking to a consistent standard. In simple terms, we want them to do one task (marking) to a high standard, not two tasks (marking and feedback) to a lower standard.

The examiner can also only make their judgment on the one piece of work they have available, and experienced teachers will draw upon a wide range of information when deciding how to offer feedback to a candidate. This means the quality of any examiner feedback will suffer from having less insight than that of the teacher.

For all these reasons, the IB is very clear to its examiners that they should only mark the candidate's work according to the correct standard and not add comments to provide feedback to the candidate or teacher.

Examiners are required to indicate clearly where marks have been awarded, and, if there could be ambiguity, to clarify with appropriate comments. This supports the IB in checking standards and also provides transparency for schools on where marks have been awarded.

The exception to this principle of only commenting where it supports the marking is where a school has requested a category 1 EUR report. In this case, a senior examiner will address the specific concerns that a school has raised when requesting the EUR, which will go beyond the usual level of detail we expect of examiners.

Subject reports

Every CE is required to produce a subject report after each session. The purpose of this report is to provide teachers with information about how the entire candidate cohort performed in this session, including

questions and topics that were addressed particularly strongly or poorly. The report provides details on each component and discusses the overall quality of candidates' answers and any general recommendations on how they could be improved for future sessions. Many subject reports also provide headline data from teacher feedback on the examination which was considered during grade award. Finally, these subject reports also provide the grade boundaries for the subject and the component boundaries that contributed to mark the overall result.

Individual teachers will naturally need to put this feedback in the context of their own classes. For example, while a particular question may have been poorly answered in general, it is quite possible that all their candidates obtained high marks.

Internal assessment (IA) feedback

The purpose of IA moderation is to ensure that all teachers are marking to the same standard, and ultimately, the IB would like no moderation factors to be applied to any school's work. To help teachers understand how they are varying from the global standard, the IB provides feedback on the IA marking so that teachers can understand why a moderation factor has been applied.

The feedback provided to schools is not intended to explain how the candidates in the sample could have achieved a better result.

Note on textbooks, workshops and examinations

The curriculum of each IB course is set out in the subject guides and this is the basis on which assessments are designed. Any textbooks, including those endorsed by the IB, are intended as aids to support teachers and learners in completing the course as set out in the subject guides and are not written to define the scope of the curriculum.

Therefore, if a topic, or part of a topic, is not included in a particular textbook but is in the guide, then questions may still be asked on it within the examinations. We would therefore encourage all teachers to refer to the subject guides exclusively when considering the scope of the curriculum. As part of the examination writing process we do check commonly used textbooks to ensure that the examination questions do not duplicate those posed in any of these resources.

Similarly, comments and handouts from IB-run workshops do not replace statements set out in the subject guides, although they may support teachers in understanding how to interpret the wording in the guide. Any amendments to the guides will be formally published by the IB and clearly described as such.

What are programme-specific processes?

- The broad purpose of having external summative assessments in IB programmes is to provide students with a "currency" to support them in progressing in education or work.
- As the same IB educational philosophy underpins all our programmes, it follows that all IB assessments should follow the same principles and broad practices to meet this philosophy. The purposes of the assessment will naturally change as the students progress through the different programmes at different stages of educational maturity.
- While assessment is an important element of the PYP, it is not appropriate for the programme to have any external summative assessment.
- Each programme has some distinctive features that require specific processes within the framework of the assessment practices.

Purpose of having external summative assessment in IB programmes

The purpose of having external summative assessment in the DP, CP and as an option for the final year in the MYP is to provide "currency" for students who have completed an IB education to support their progression to future education or employment.

The IB needs to provide this external summative assessment to ensure that the "currency" available to IB students meets the mission and philosophy of the IB.

- Valuing holistic programme-driven education rather than narrow subjects
- Valuing international-mindedness, through the learner profile
- Valuing meaningful, authentic assessment opportunities over procedural, "fact recall"-driven assessment
- Allowing for fair differentiation of students in appropriate contexts
- Encouraging a positive backwash effect on teaching and learning

These values are encapsulated in our principles of assessment.

Students taking our PYP have much less need for such a "currency" and so there is not the same need for externally assessed summative assessment. The principles of good assessment that underpin our IB philosophy still apply.

Assessment within the IB also has a secondary purpose of providing leadership through example in a world where assessment is highly valued.

Finally, summative assessment acts as confirmation for teachers, schools, parents and students that their school's interpretation of a high-quality education matches that of the IB.

Programme-specific needs and solutions

In the earlier sections, we have explained the processes that form the assessment cycle, and these generalized practices hold true across all of our programmes. Each programme serves a different purpose however and so there will also be differences in some aspects of the assessment, for example, how the extended essay or personal project are formulated and assessed and/or the rules to determine whether a programme certificate is given.

These approaches to assessments, which are specific to each programme, are important and in this section we explain what they are and set out how they are managed.

Transition from curriculum design to assessment and back to curriculum design

In the previous two sections, we explained that assessment must be integrated with the aims of the course. It follows that the assessment model must form part of the development of the course and influence the course development if assessment outcomes are going to be a meaningful and fair reflection of what the course seeks to provide.

In the IB, the approach to reviewing and revising subjects includes discussion of assessment, and part of the teacher materials we offer to schools includes samples of what future assessments will look like and examples of candidate work with notes on how they would be marked or assessed.

The purposes of sample examination material, such as specimen papers, are:

- to indicate the structure of the examinations, that is, where candidates must answer all questions in a section, where there is optionality, and the length and proportion of marks for each section and paper
- to provide an indication of the style of questions and types of stimulus material in the examination, particularly for elements of the course that are new
- to provide indicative content for a live examination paper, particularly for elements of the syllabus outline (or of the curriculum) that are prescribed versus optional
- to provide teachers with material to **reasonably** prepare candidates for their examinations, for example, through a "mock" examination.

The specimen papers can provide example questions on all aspects of the course, or even the new aspects of a course. The specimen papers should be designed in such a way that they could be reasonably used as a "real" examination by the IB, not developed purely as specimen material.

Elements common to all programmes

- IB programmes offer curriculum or curriculum frameworks that are broad, balanced, conceptual and connected.
- All IB assessments need to consider these underlying aspects of an IB education in their design, even when they are not explicitly assessed, so that there is a positive backwash effect on teaching and learning.
- The key elements that link all IB programmes are:
  - the learner profile
  - approaches to teaching and learning
  - international-mindedness.

Figure 63

Continuum between IB programmes

In *What is an IB education?* we emphasize that our programmes promote conceptual learning, focusing on powerful organizing ideas that are relevant across subject areas, and that help to integrate learning and add coherence to the curriculum. The programmes emphasize the importance of making connections, exploring the relationships between academic disciplines, and learning about the world in ways that reach beyond the scope of individual subjects. They offer students access to a range of academic studies and learning experiences which are broad, balanced, conceptual and connected.

- Broad, balanced—An IB education represents a balanced approach, offering students access to a broad range of content that spans academic subjects.
- Conceptual—Conceptual learning focuses on broad and powerful organizing ideas that have relevance within and across subject areas. They reach beyond national and cultural boundaries. Concepts help to integrate learning, add coherence to the curriculum, deepen disciplinary understanding, build the capacity to engage with complex ideas and allow transfer of learning to new contexts.
- Connected—IB curriculum frameworks value concurrency of learning. Students encounter many subjects simultaneously throughout their programmes of study; they learn to draw connections and

pursue rich understandings about the interrelationship of knowledge and experience across many fields. Course aims and programme requirements offer authentic opportunities to learn about the world in ways that reach beyond the scope of individual subjects.

As part of this, all programmes include a culminating project in their assessment. In the PYP, this is the exhibition, for the MYP the personal project or community project, in the DP the extended essay, and in the CP the reflective project.

Even when designing individual assessments, teachers, schools and IB authors need to reflect upon these underlying goals in order to avoid creating tasks that undermine good teaching and learning. The aim must always be to generate a positive backwash effect.

In delivering the IB mission, good quality programmes encompass three areas:

1. the learner profile
2. approaches to teaching and approaches to learning
3. international-mindedness and intercultural understanding.

In meeting the broader objectives of these programmes, and thus being fit for purpose (valid), IB assessments need also to include consideration of these elements, even when they are not what is intended to be assessed.

Student competencies and the learner profile

> Education today is much more about ways of thinking which involve creative and critical approaches to problem-solving and decision-making. It is also about ways of working, including communication and collaboration, as well as the tools they require, such as the capacity to recognize and exploit the potential of new technologies, or indeed, to avert their risks. And last but not least, education is about the capacity to live in a multi-faceted world as an active and engaged citizen. These citizens influence what they want to learn and how they want to learn it, and it is this that shapes the role of educators.
>
> (Andreas Schleicher 2016)

It is increasingly being claimed that the skills that are required this century are fundamentally different to those of previous generations. While there are those who would argue that the inquiry approach that underpins these 21st century skills has been valued since Socrates, there is general agreement of the importance to provide students with a wide range of attributes to prepare them for life. See, for example, the arguments made in Llewellyn (2014).

There is a wider range of different ways of categorizing these skills, including OECD's 21st century competencies, RAND Education, NRC Framework and others. Within the IB, we describe these competencies within the learner profile.

Figure 29

IB learner profile

Not all aspects of the learner profile are appropriate to measure through summative assessment, but several are encapsulated within the concept of higher-order thinking skills. Good assessment recognizes the importance of these characteristics and even when it is not designed to measure them, it can offer students a chance to develop these competencies. Examples of this could be through encouraging ethical (principled) approaches to surveys and experimentation, supporting appropriate peer review and introducing unexpected contexts to students.

For more details on the IB's wider approach to student competencies refer to the material available on the IB website or the relevant programme's *From principles into practice* document.

Approaches to teaching and approaches to learning

The IB aims not to be prescriptive or restrictive in its approaches to learning (ATL) and approaches to teaching, but to focus on offering guidance and suggestions. We recognize the need to allow individual teachers and schools to have space to be creative, although there is a need to focus discussion and highlight good practice. We do believe that in order for the teaching of skills to be effective, ATL need to be both explicitly articulated and sustained in their implementation.

Improving skills requires reinforcement over an extended period of time, and in a variety of contexts. Whatever approach teachers or schools decide to use to embed ATL in their classroom is a decision the IB believes should be left to the people with the deepest insight into the needs of their students, that is, the teachers.

For more details of the IB's approach please refer to the *Approaches to teaching and learning* resources, or relevant sections of the programme's *From principles into practice*.

International-mindedness and intercultural understanding

IB programmes are studied by students in many countries and of many nationalities. As well as the academic aims of our programmes, the IB intends that students should develop as "caring young people who help to create a better and more peaceful world through intercultural understanding and respect", and "who understand that other people, with their differences, can also be right" (IB mission statement 2002). There is, therefore, both an international context and an intercultural understanding purpose to IB teaching, both of which must be reflected in the assessment.

The most important step in delivering this is through having academic experts, including examination paper authors and curriculum developers, from a wide range of cultural backgrounds. It is important to the IB mission not to obscure differences but to engage with them in a way that allows students to explore them without being disadvantaged.

In some subject areas, the issue of cultural variety can be encouraged through a recognition of different cultural emphasis in the curriculum. Examples of this approach can be found in biology, chemistry, psychology and visual arts. In the first three of these, the option structures within each subject allow schools to select course content that will, to a certain extent, suit particular cultural traditions of teaching the subject.

In other subject areas, international-mindedness is encouraged through the material and inspiration the student is encouraged to use. Examples of this includes the arts subjects and literature and language but it also can be included through a wider range of internal assessment tasks.

There is more to international-mindedness than just knowledge and understanding of other cultures. Attitude and action are also important attributes. Attitudes are difficult to assess through normal school assessment, which focuses on achievement rather than affective attributes.

Within the IB programmes, this is addressed through the non-assessed elements of the course such as the creativity, activity, service (CAS) part of the DP and the community project in MYP. As the diploma and the MYP certificate cannot be awarded without candidates having completed this aspect, these non-assessed elements have a significant impact on the overall outcome of IB assessment.

Figure 30

Principled action

While allowing candidates to choose which questions to answer might be seen as the best way of addressing the different international requirements in assessment, this then poses assessment problems in terms of maintaining comparability across the options. This always occurs when there are choices of question, or very open-ended assessment tasks. It is challenging to even define what "equal demand" means when the candidates come from very different educational backgrounds. In general, it is easier to maintain comparability by setting common tasks which allow candidates to introduce their own experiences into the answers. In such cases, the challenge falls upon the examiner to maintain a common standard, but this is one step easier than having two separate tasks of potentially different levels of demand which must then also be marked to the same standard.

Information on comparability can be gained through analysis of candidate performance and this analysis is discussed further in the section on "Grade awarding (aggregation)".

Assessment carried out in an international context has additional challenges in terms of equity, above those normally encountered within a national system. Questions that might be perfectly appropriate in one national setting become inappropriate in another. Questions referring to sports, travel, entertainment, historical events, even the weather, must be prepared very carefully. It might seem that the only way around this problem is to prepare examination questions that are devoid of all but a lowest common denominator of sociocultural context. However, to do so would not only make examination questions very limited and dull, it would also be against the whole philosophy of IB assessment and against good assessment practice in terms of ensuring validity through context-based tasks. Contextualized work and assessment are vital to good learning.

There are two possible ways around this dilemma. First, background contextual information can be provided to candidates, through specification in the subject syllabus content, by providing case studies on which questions are based, or even in the examination question itself (as long as this is not too lengthy and thus distracting from the purpose of the assessment).

A second method is to use more open-ended assessment questions and tasks that allow candidates to select their own context in which to respond. In the latter approach, the focus of marking must be on deeper levels of understanding, rather than on straightforward knowledge of subject content, since there will be no common basis of content. This is very much in keeping with the IB assessment philosophy.

Figure 31

Range of cultural norms/contexts

Even with the application of both these methods, candidates may find themselves dealing with assessment tasks having contexts that are not familiar to them within their own sociocultural background. This again is in keeping with our assessment philosophy, in that one of the aims of the programmes is to make students more open-minded to other ways of doing things, more globally aware, and more competent at operating in a non-familiar cultural environment. Part of the requirement for higher-order thinking is that students should be able to apply knowledge in unfamiliar situations. It is quite appropriate for such elements to be included in assessment, as long as they affect students from different cultural backgrounds evenly.

A significant proportion of IB students enter for examinations in a language that is not their best. Nearly all such cases relate to English, because students working in French or Spanish (the other two main languages in which IB assessment is conducted) tend to be native speakers. Considerable extra care has to be taken in the wording of questions so as not to disadvantage second-language speakers. This is dealt with in paper editing.

Our summative assessment, along with the great majority of formal assessment systems, is highly individualistic. As pointed out by Brown (2002), this is largely because the DP falls within the western European tradition, and western European societies are individualistic in nature. Candidates are assessed almost exclusively on what they achieve on their own. This may be said to be culturally inequitable, since there are a number of cultures in which the contribution of the individual is always subservient to that of a larger group; it is what the group achieves that matters. It is also the case that, in terms of individual equity, there are some people who work better in a team than they do individually, and vice versa. Additionally, it is common practice, both in the classroom and in the world of work, for individuals to work interdependently rather than independently.

Further reading

For more information about the values that underpin the wider IB educational programme please refer to the following resources.

- *What is an IB education?*
- *Approaches to teaching and learning*
- The IB learner profile
- Individual programme principles into practice guides—MYP, DP, CP

IB Diploma Programme

The distinctive features of DP external summative assessment are:

- students must take a prescribed set of subjects to achieve the diploma
- achievement in the overall diploma is described by a points score whose maximum is 45
- core subjects contribute up to three points to overall diploma outcome via a points matrix
- nearly all subjects have multiple components which cover both external and internal assessment
- nearly all subjects are available at standard level (SL) or higher level (HL), and contribute equally to the overall diploma outcome
- subjects differ considerably in the number of candidates taking them.

Aims of the Diploma Programme

The validity of assessment outcomes can only be determined if we are clear what the purpose of the course and programme are. For this reason, we start this section by discussing the aims of the programme.

> The Diploma Programme (DP) provides a challenging, internationally focused, broad and balanced educational experience for students aged 16 to 19. Students are required to study six subjects and a curriculum core concurrently over two years. The programme is designed to equip students with the basic academic skills needed for university study, further education and their chosen profession. Additionally, the programme supports the development of the values and life skills needed to live a fulfilled and purposeful life.
>
> (*Diploma Programme: From principles into practice* 2010: 15)

What makes the Diploma Programme (and indeed all IB programmes) special is that we are concerned with the whole educational experience of each student, and this is reflected in the focus on programme level validity, not just each individual course.

Valid uses for outcomes of diploma assessments

When developing assessment models and curriculum we intend that grades from DP courses and the overall diploma points score should be used to determine:

- selection for university admission or work
- whether students have already met the requirements of a university programme (either additional credit or excused from taking particular studies/courses).

Where a candidate has taken the assessment in a particular (response) language, that also provides evidence that they can undertake further study in that subject in that language.

See the section on use of qualifications for why these valid uses are important considerations.

Structure of the DP

In order to achieve the IB diploma certificate, a candidate must take six subjects, together with the core elements—theory of knowledge (TOK), the extended essay (EE), and creativity, activity, service (CAS).

Students choose courses from the following subject groups:

- studies in language and literature
- language acquisition
- individuals and societies

- sciences
- mathematics
- the arts.

Students may opt to study an additional sciences, individuals and societies, or languages course, instead of a course in the arts.

Students will take some subjects at standard level (SL) and some at higher level (HL). SL and HL courses differ in scope but are measured according to the same grade descriptors, with students expected to demonstrate a greater body of knowledge, understanding and skills at higher level.

Each student takes at least three (but not more than four) subjects at higher level, and the remaining subjects at standard level.

A limited number of interdisciplinary courses count across subject groups, for example, environmental systems and societies simultaneously satisfies the individuals and societies and sciences groups. The interdisciplinary courses provide flexibility to students in choosing which six subjects to take.

Figure 64

Diploma Programme model—describing the structure of the programme

There are additional rules in order to prevent students from taking subjects whose content overlaps. These are detailed in the *Assessment procedures* for the DP.

How the diploma outcome is calculated

The overall diploma points are calculated by adding together the grades (1 up to 7) achieved from each of the six subjects and then including between zero and three points from the core. This means that the highest score that a candidate can achieve is 45 points*.

This approach means that SL and HL subjects are valued equally in determining the candidate's final points.

*The maximum points of 45 is obtained from 6 (subjects) times 7 (top grade) plus 3 points from the core.

Core points matrix

Unlike the other subjects, TOK and the EE are graded from A to E. The third element of the core, CAS, does not receive a grade as it would not be meaningful to evaluate performance in this area.

The core is worth between zero and three points towards the overall diploma points. The candidate can also fail to achieve the diploma certificate if they obtain a grade E in either TOK or EE or if they do not complete CAS. The number of points is calculated using the table below.

Figure 65

Core points matrix

		Theory of knowledge (TOK)				
	Grade awarded	A	B	C	D	E
Extended essay	A	3	3	2	2	Failing condition
	B	3	2	2	1	
	C	2	2	1	0	
	D	2	1	0	0	
	E	Failing condition				

Failure conditions

A candidate can only receive the overall diploma certificate if none of the following nine conditions below applies.

- CAS requirements have not been met.
- Candidate's total points are fewer than 24.
- An N (no grade awarded) has been given for TOK, EE or for a contributing subject.
- A grade E has been awarded for one or both of TOK and the EE
- There is a grade 1 awarded in a subject/level.
- Grade 2 has been awarded three or more times (HL or SL).
- Grade 3 or below has been awarded four or more times (HL or SL).
- Candidate has gained fewer than 12 points on HL subjects (for candidates who register for four HL subjects, the three highest grades count).
- Candidate has gained fewer than 9 points on SL subjects (candidates who register for two SL subjects must gain at least 5 points at SL).

Bilingual diplomas

As an alternative to the standard diploma certificate, a "bilingual diploma certificate" can be awarded to a candidate who:

- completes two languages selected from group 1 with the award of a grade 3 or higher in both
- completes one of the subjects from group 3 or group 4 in a language that is not the same as the candidate's nominated group 1 language. The candidate must attain a grade 3 or higher in both the group 1 language and the subject from group 3 or 4.

Pilot subjects and interdisciplinary subjects can contribute to the award of a bilingual diploma certificate, provided the above conditions are met.

The following cannot contribute to the award of a bilingual diploma certificate:

- an extended essay
- a school-based syllabus
- a subject taken by a candidate in addition to the six subjects for the diploma certificate ("additional subjects").

Approach to missing marks

There are some circumstances where we are unable to obtain or use the work that the candidate has submitted. Examples of such circumstances might be where work has been lost or destroyed in the post or there is evidence of wide-scale academic misconduct but we cannot identify which work may have been affected. In such circumstances, the IB may attempt to estimate the grade they would have obtained through the missing mark process.

The high-level process and its strengths and weaknesses are described in the "Missing marks" section. Below we describe the calculation we use in the DP.

If there are five or more candidates for the subject in a school, we use the school's data to determine missing marks. If there are fewer than five candidates in a school, then we use global data to calculate missing marks. Our experience suggests that there is usually little difference in outcome when using school or global data when the school has five or more candidates.

Figure 66

How to determine a candidate's missing mark

To determine the candidate's missing mark:

$$Ratio = \frac{\text{Candidate's total scaled mark on other components}}{\text{Global (school) total scaled mark on other components}}$$

$$\text{Candidate's missing mark} = \text{global (school) scaled mark on missing components} \times \text{ratio}$$

The ratio compares the candidate's marks with the global or school average for the components they have completed:

- If the candidate has done better than average this number will be greater than one.
- If the candidate has done worse than average this number will be less than one.

This calculation, like all missing mark procedures, is only the "best guess" based on the available evidence. It is always preferable to have actual candidates' work to mark, and if necessary apply our special consideration processes.

Managing standards across the range of sizes of cohorts

Diploma subjects can range in size from a single candidate to over 40,000. While every subject is equally important, and candidates will make the same use of their grades whether they are the only candidate or one of thousands, for practical reasons we have slightly different processes for subjects with large and small candidate numbers.

Maintaining examiner quality

The quality model of practice, qualification and seed scripts is designed to ensure all examiners have understood and are applying the PE's standard. If there are very few examiners, then this model may not be appropriate to use.

If all examiners are part of the discussion at standardization, then it is not necessary to provide practice scripts or qualification scripts, as every examiner has already been part of setting the standard. Seed scripts are still used to ensure that examiners continue to mark to the expected standard.

Since there is always the risk that we will need to introduce additional examiners if marking is slow, we generally produce a complete quality model even if it may not end up needing to be used. Even if we have not created qualification scripts, we can use seed scripts for qualification purposes if necessary, although this is not ideal as there are different considerations when selecting the two different types of definitively marked scripts.

The critical factor is that marking quality is essential even when there is a small number of candidates in a subject and the IB makes use of all available evidence to identify any issues before issue of results.

Grades—Full grade award (virtual or face-to-face)

For subjects where there are several examiners under the PE, we hold a formal meeting to follow the practices set out in the "Grade awarding (and aggregation)" section. An IB subject manager will be involved in the meeting to support the CE and to provide a quality control check on the process.

These meetings may be either face-to-face or virtual. In the latter case they need to be managed to reflect the fact that senior examiners can be in different time zones around the world. In such meetings, the senior examiners will use a private online forum to share their thoughts as they work through the evidence as well as having video conference sessions to agree key decisions.

Grades—Guided small entry subject grade award (virtual)

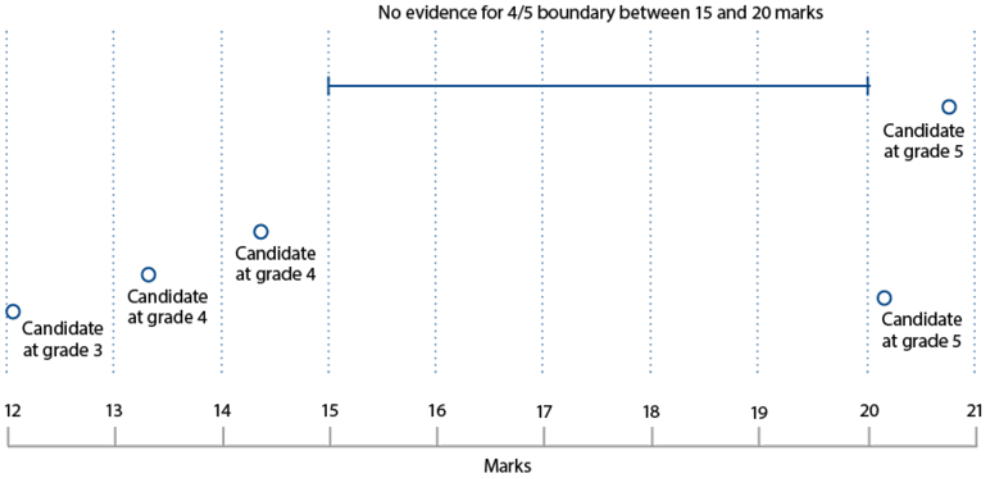
Where there is a very limited number of examiners in each component, the senior examiners will undertake the discussion virtually. The IB subject manager will provide them with the full range of evidence to support the award and monitor the progress to ensure quality but is unlikely to be present for the discussions. The senior teams for such small entry subject grade awards are generally themselves very small.

Grades—Standard small entry grade awards

When the number of candidates is small enough that the statistical evidence between years is much less meaningful, we ask our examiners to review all candidate work rather than just a sample at the judgmental grade boundaries. In these circumstances, the PEs and CEs can suggest each grade boundary rather than just the 2/3, 3/4 and 6/7. This is because there can often be no evidence of candidate work at a particular grade boundary. Consider the example illustrated below:

Figure 67

Example of situations where there is a lack of evidence at a grade boundary



The IB subject managers support these meetings as requested and carry out random quality checks rather than routinely observing the meeting. These small entry subjects are some of the most challenging to maintain a consistent standard in as the small number of candidates limits the value of historical data. These candidates are also frequently bunched at the top of the mark range.

To support examiners in all small entry subjects (guided and standard grade award) we encourage discussion between examiners in different subjects but within the same group. This helps to ensure there is a common understanding of what grades mean across the group.

Final sign off

Irrespective of the number of candidates who have taken the examinations, the final stage in the grade award process is for the CE to make recommendations to the IB and for senior assessment staff to confirm they are convinced by the evidence supporting these recommendations.

While we do take into account the relative meaningfulness of the statistical evidence based on the number of candidates in the cohort, every subject is scrutinized in the same way.

Further reading

For more information about the IB Diploma Programme please refer to the following resources.

- *Diploma Programme: From principles into practice*
- *General regulations: Diploma Programme*
- Diploma Programme *Assessment procedures*
- *Rules for IB World Schools: Diploma Programme*
- Subject guides (see subject pages on the programme resource centre)
- Teacher support materials (see subject pages on the programme resource centre)

IB Career-related Programme

The distinctive features of CP assessment are:

- students must meet a set of requirements to achieve the CP
- there is no overall points score associated with the CP certificate
- CP students taking courses shared with DP are assessed jointly with diploma students
- the CP framework requires a career-related study which is not offered or awarded by the IB.

Aims of the Career-related Programme

The validity of assessment outcomes can only be determined if we are clear what the purpose of the course and programme are. For this reason, we start this section by discussing the aims of the programme.

The unique feature of the CP is that it supports students to become career-ready learners, in whatever career they have chosen. The course ensures that they develop the transferable and lifelong skills to support them throughout their employment however they choose to progress.

The programme helps students to:

- develop a range of broad work-related competencies and deepen their understanding in specific areas of knowledge through their Diploma Programme courses
- develop flexible strategies for knowledge acquisition and enhancement in varied contexts
- prepare for effective participation in the changing world of work
- foster attitudes and habits of mind that allow them to become lifelong learners willing to consider new perspectives
- become involved in learning that develops their capacity and will to make a positive difference.

(*Career-related Programme: From principles into practice* 2015: 7)

Valid uses for outcomes of CP assessments

When developing assessment models and curriculum, we intend that CP course grades and the certificate can be used to determine:

- selection for employment and employment programmes such as apprenticeships
- selection for further education in the appropriate vocational field of study
- selection for university
- whether students have already met the requirements of a university programme (either additional credit or excused from taking particular studies/courses).

Where a candidate has taken the assessment in a particular (response) language that also provides evidence, they can undertake further study in that subject or vocation in that language.

See the section on use of qualifications for why this is an important consideration.

Structure of the CP

The CP is a three-part educational framework. It consists of:

- at least two courses from the DP at standard level (SL) or higher level (HL)
- the CP core
- a career-related study.

Figure 68

Career-related Programme model—describing the structure of the programme

The core subjects including the reflective project

The core subjects are intended to contextualize both the DP courses and the career-related study. It is intended to act as the conduit to link all the areas of learning together.

Completing the four elements of the core is mandatory. They are:

- personal and professional skills
- service learning
- language development
- reflective project.

The IB only assesses the reflective project, which is teacher marked and then moderated by the IB. For the other elements, the school must confirm to the IB that they have been completed satisfactorily but the IB does not verify the assessment (if any).

Diploma courses as part of the CP

Each student takes at least two subject courses which are common with the DP. These CP candidates are included in the same assessment process as the DP candidates. There is no separate examination or grade award for CP.

A candidate cannot be registered simultaneously for completing the DP and the CP, despite courses being common to both. The extent of the wider programme requirements of each preclude them being completed simultaneously.

Career-related study

The IB does not assess or apply any sort of quality-control to the outcomes of the career-related study portion of the CP. The only requirement is that the school confirms that the student has completed it. IB diploma courses are not appropriate to form part of the career-related study portion of the CP.

How the CP outcome is calculated

There is no points score associated with the CP certificate.

The CP certificate will be awarded to a candidate provided all of the following requirements have been met.

- The school has confirmed that the candidate has completed the specified career-related study.
- The candidate has been awarded a grade 3 or more in at least two DP courses.
- The candidate has been awarded at least a D grade for the reflective project.
- The school has confirmed that all personal and professional skills, service learning and language development requirements have been met.
- The candidate has not received a penalty for academic misconduct from the final award committee.

The career-related diplomas and reflective project grades are confirmed by the same final award committee as the DP.

Bilingual CP certificates

In addition to the usual certificate, a "bilingual certificate" can be awarded to a candidate who:

- completes two DP language courses selected from studies in language and literature with the award of a grade 3 or higher in both
- completes a DP language course from studies in language and literature and also completes a DP course from individuals and societies or sciences in a response language that is not the same as that taken from studies in language and literature. The candidate must attain a grade 3 or higher in both courses.

Managing standards and missing work

The reflective project and all courses that are common with the DP use the same approaches to the DP for missing mark procedure and managing standards across the range of sizes of cohorts.

Further reading

For more information about the IB Career-related Programme please refer to the following resources.

- *Career-related Programme: From principles into practice*
- Career-related Programme *Assessment procedures*
- *Overview of the Career-related Programme*
- *Reflective project guide*
- *Language development guide*
- *Personal and professional skills guide*
- *Service learning guide*

IB Middle Years Programme

The distinctive features of MYP assessment are:

- all candidates in year 5 must complete an externally moderated personal project, but other IB assessments are optional for schools teaching the MYP
- students must take a prescribed set of subjects to achieve the MYP certificate
- achievement in the overall MYP certificate is described by a point score whose maximum is 56
- the core subjects of interdisciplinary learning and personal project contribute equally with the other subject disciplines. Community service does not contribute to the total
- for the IB-designed summative assessments:
  - each subject has only one component
  - subjects are assessed either by e-portfolio or on-screen examination.

Aims of the Middle Years Programme

The validity of assessment outcomes can only be determined if we are clear what the purpose of the course and programme are. For this reason, we start this section by discussing the aims of the programme.

> The MYP has been designed as a coherent and comprehensive curriculum framework that provides academic challenge and develops the life skills of students from the ages of 11 to 16. These years are a critical period in the development of young people. Success in school is closely related to personal, social and emotional well-being. At a time when students are establishing their identity and building their self-esteem, the MYP can motivate students and help them to achieve success in school and in life beyond the classroom. The programme allows students to build on their personal strengths and to embrace challenges in subjects in which they might not excel. The MYP offers students opportunities to develop their potential, to explore their own learning preferences, to take appropriate risks, and to reflect on, and develop, a strong sense of personal identity.
>
> (*MYP: From principles into practice* 2014: 3)

The MYP has an explicit alignment between the MYP subject group objectives and marking criteria. All MYP subject groups have four assessment criteria which match the four objectives. Each criterion contributes equally to the final outcome.

Inclusion of the global context in eAssessments

> In the MYP, learning contexts should be (or should model) authentic world settings, events and circumstances. Contexts for learning in the MYP are chosen from global contexts to encourage international-mindedness and global engagement within the programme … The MYP identifies six global contexts for teaching and learning that are developed from, and extend, the PYP's transdisciplinary themes.
>
> (*MYP: From principles into practice* 2014: 18)

Each examination session will be shaped within and informed by a specific global context and exploration selected from the list published in *MYP: From principles into practice*.

Approximately one third of tasks within each disciplinary on-screen examination will be connected with, inspired by or derived from the selected global context. The whole of the interdisciplinary learning on-screen examination is inspired by the selected global context.

Partially completed unit planners for language acquisition, arts, design, and physical and health education will be developed with reference to the selected global context.

Valid uses for outcomes of MYP assessments

When developing assessment models and curriculum, we intend that grades from MYP courses and the certificate points score should be used for:

- selection for further educational opportunities or work
- positive feedback and an indication of personal strengths for students continuing their education.

Where a candidate has taken the assessment in a particular (response) language, that provides evidence that they can undertake further study in that subject in that language, for example, studying at a French, Spanish or English school.

See the section on use of qualifications for why this is an important consideration.

Structure of the MYP

Figure 69

Middle Years Programme model—describing the structure of the programme

How the MYP outcome is calculated

At the end of their year 5 MYP studies, candidates can be entered for the IB external assessment. The outcomes of these assessments will be recorded in an MYP Course Results document. In addition, these candidates can choose to take assessments which can lead to the award of the MYP Certificate.

The school can also issue an MYP Record of Participation. This is for MYP students who study the programme for at least two years and complete the requirements in year 3 or year 4. These students are not registered with the IB for any form of assessment. The Record of Participation is a school-based document, not verified by the IB.

In order to achieve the IB MYP Certificate, the student must have participated in the final year of the programme, with a recommended period of participation of two years, and:

- complete either an on-screen assessment or ePortfolio in six subjects consisting of: language and literature, language acquisition (or a second language and literature), individuals and societies, mathematics, sciences and one subject from arts, physical and health education or design
- achieve at least a grade 3 in each of the six subjects above
- complete the on-screen examination in interdisciplinary assessment and achieve at least a grade 3
- complete the personal project with at least a grade 3
- obtain a total of 28 points overall
- meet the school's expectations for community service.

The MYP bilingual certificate additionally requires successful results from on-screen examinations for one of the following:

- a second language and literature course (instead of a course in language acquisition)
- one (or more) science, individual and societies, or interdisciplinary examination in a language other than the student's chosen language and literature course.

Delivering external summative assessment (MYP eAssessment)

The optional eAssessment comprises two different ways to assess what students know and can do:

- ePortfolios of candidate work in language acquisition, arts, design, and physical and health education, which are then moderated to ensure a consistent global standard
- on-screen examinations (two hours in duration) for courses in language and literature, individuals and societies, sciences, mathematics, and interdisciplinary learning.

In addition, the personal project is submitted electronically to the IB and moderated. While other eAssessments are optional for schools, all MYP year 5 students must take part in the personal project eAssessment.

Examination blueprints

The IB publishes examination blueprints to provide clear guidance to schools on what the eAssessments will look like. These blueprints enable teachers and candidates to understand the nature and purpose of MYP eAssessment. They assist candidates to prepare for on-screen examinations, and help candidates to focus on the subject-group criteria and assessment strategies in each subject group. There are always four criteria in the blueprint and each of these criteria is equally weighted.

The IB undertakes to ensure that in any session, examinations will not deviate from the blueprint by more than three marks.

ePortfolios and partially completed unit planners

ePortfolios allow the assessment of an extended coursework task (product) or performance which by their very natures are difficult to test through an examination. The basis of the ePortfolios are the partially completed unit planners which guide the teacher in ensuring that appropriate candidate evidence is produced to allow fair and meaningful judgments to be made as well as providing flexibility to meet local contexts. New partially completed unit planners are provided for each session.

The unit planners should ensure that the tasks set by teachers allow candidates to show evidence across the full range of MYP grades. There is clearly a risk with teacher-devised assessment that candidates are disadvantaged by an unreasonably easy or difficult set of tasks. In moderation, the IB can only award grades based on the candidate work available and if the teacher-devised task only covers part of the range of grade descriptors we cannot award grades outside that range.

Single assessment—managing candidate burden

Taking examinations and doing coursework is stressful and demanding on candidates. It can also take away from time spent teaching. For the 16-year-olds studying the MYP, the IB believes that, on balance, it is more appropriate to minimize the amount of summative assessment. While this does create difficulties with candidates only having one opportunity to demonstrate what they can achieve, the IB accepts these problems to ensure the overall welfare of the candidates.

Approach to missing grades

One disadvantage of managing candidate burden by only having one assessment component is that there is only limited evidence available in the MYP. This means that our confidence in the missing grade procedure process is lower than for other programmes and therefore should only be used in exceptional cases.

- Where a candidate has not undertaken an assessment they should be offered the opportunity to take the exam in a later session.
- Where a candidate has undertaken the assessment but external factors are likely to have affected their performance they should receive special considerations.
- Where a candidate has completed the assessment as required and submitted it in good faith but it is not available for the IB marking due to factors beyond the control of the candidate or school then we will apply the missing grade procedure.

In order to apply the missing mark procedure the candidate must have been awarded a final grade in at least four other courses with the MYP. We are unable to estimate results for candidates with fewer results because we do not have enough evidence to make an informed estimation.

The missing grade calculation involves determining the mean (average) grade from all other subjects with a grade determined by candidate work in the past 18 months.

- If the mean grade is 0.5 or higher—round up.
- If the mean grade is less than 0.5—round down.

This calculation, like all missing mark procedures, is only the "best guess" based on the available evidence. It is always preferable to have actual candidates' work to mark, and if necessary, apply our special consideration processes.

Further reading

For more information about the IB Middle Years Programme please refer to the following resources.

- *MYP: From principles into practice*
- *Guide to the MYP exam session*
- MYP subject guides
- *MYP projects guide*
- Middle Years Programme *Assessment procedures*
- *IT requirements for conducting MYP on-screen examinations*
- *MYP on-screen familiarization for students* (PC), (MAC)

IB Primary Years Programme

The distinctive features of PYP assessment are:

- Assessment involves teachers and students collaborating to monitor, document, measure, report and adjust learning
- there is no requirement for IB external summative assessment.

Aims of the Primary Years Programme

The validity of assessment outcomes can only be determined if we are clear what the purpose of the programme is. For this reason, we start this section by discussing the aims of the programme.

> The PYP focuses on the heart as well as the mind and addresses social, physical, emotional and cultural needs in addition to those considered to be more academic. The traditional subject areas are valued, with an extra emphasis on the balance between the acquisition of essential knowledge and skills and the search for the meaning of, and understanding about, the world. The programme provides the opportunity for learners to construct meaning, principally through concept-driven inquiry. The threads of students' learning are brought together in the transdisciplinary programme of inquiry, which in turn allows them to make connections with life outside the school.
>
> (IB PYP documentation)

Structure of the PYP

Figure 70

Primary Years Programme model—describing the structure of the programme

Assessment—formative support for learning

The prime objective of assessment in the PYP is to provide feedback on the learning process. Bruner states that students should receive feedback "not as a reward or punishment, but as information" (Bruner 1961: 26). Teachers need to select assessment strategies and design assessment instruments to reflect clearly the particular learning outcomes on which they intend to report. They need to employ a range of strategies for assessing student work that takes into account the diverse, complicated and sophisticated ways that individual students use to understand their experiences. Additionally, the PYP stresses the importance of both student and teacher self-assessment and reflection.

Further reading

For more information about the IB Primary Years Programme please refer to the PYP resources page on the programme resource centre.

Annex 1: Moderation of internal assessment

Moderation is used with internally assessed work to ensure a common standard across all schools. As a result of moderation, a school's marks may be lowered, raised or remain the same. The aim of moderation is to check how accurately and consistently the teacher has applied the assessment criteria in his or her marking of the candidates' work.

Sampling

- Teachers within a school must mark to a common standard as, where necessary, one moderation factor is applied to all candidates in a subject.
- Candidates who have attained full marks tend not to be selected for the sample to allow higher marks the possibility of being moderated upwards.

The internal assessment (IA) sample is carefully selected to ensure that the mark range of the school is appropriately represented. Moderation sample sizes are ten, eight, five, or fewer than five, according to the number of candidates in the subject cohort. The IB selects the candidates whose work comprises the internal assessment moderation samples after the school has formally submitted its internal assessment marks. The IB tends not to select candidates for the moderation sample who have attained full marks, to allow candidates in the higher mark range the possibility of being moderated upwards.

When the school entry for a given course is large enough to split into different classes and more than one teacher is involved in carrying out the internal assessment, the IB requires these teachers to share the internal assessment and work together to ensure they have standardized between them the way in which they apply the criteria. A single moderation sample is requested from the school, which in all probability will contain candidate work marked by the different teachers involved. However, where there are different classes within one school using different response languages for the same subject, then a separate moderation sample is required for each language.

Determining moderation factors

All internally assessed components are marked by applying assessment criteria or markbands, and in the majority of cases the teacher has access to considerably more information about the context and process underlying the candidate work than the examiner can have. Because of this, examiners moderating internally assessed components are asked to judge whether the teacher's marking seems appropriate, rather than simply to re-mark the work disregarding the marks awarded by the teacher. Teachers' marks should be altered only when the moderator is sure they are inappropriate.

The teacher's marking sample is moderated by the examiner and, based on a statistical comparison between the two sets of marks (using linear regression), an adjustment is made to the teacher's marks for all candidates at the school for that component.

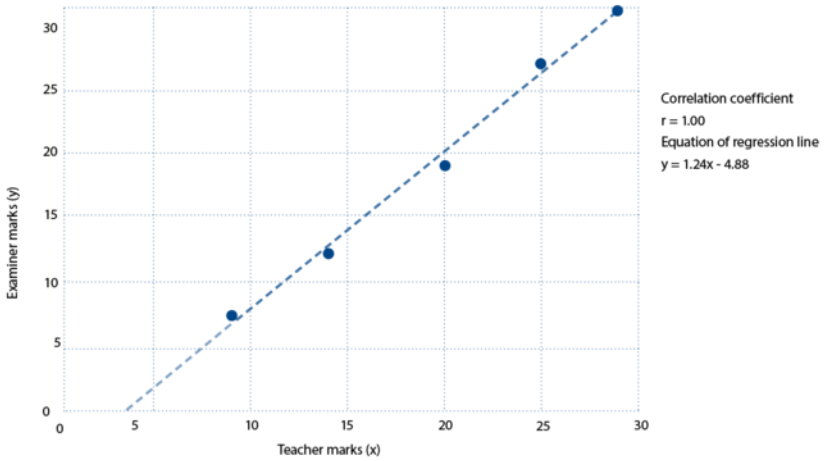
- If the teacher is consistently under- or over-marking, this adjustment will be the same for each of the teacher's marks.
- If the teacher is under- or over-marking either at the top or bottom of the mark range, the adjustment may vary across the range of the marks.
- If the teacher is marking to the correct standard no adjustment will be made.

Linear regression

An analysis is carried out on the data for each moderation sample, which permits an appropriate adjustment to be applied to all of a teacher's marking based on the general trend shown in the sample. The technique used is called linear regression, which involves calculating the best-fitting straight line through the set of data points derived from the sample marks awarded by both the teacher and the examiner. An example linear regression and the corresponding set of marks are shown below.

Figure AN1

An example linear regression and the corresponding set of marks



The moderation regression line for a teacher who is slightly too harsh at the top end and too generous at the bottom. Each individual point represents the pair of marks given to a piece of sample work by the teacher and the examiner. The continuous regression line is used to convert the teacher's marks into moderated marks.

Figure AN2

The corresponding marks awarded after the teacher marks and examiner marks are compared using linear regression. The calculated line of best fit, or "moderation factor", gives the final moderated marks.

Teacher mark	Examiner mark	Final moderated mark
28	30	30
25	27	26
20	19	20
14	12	12
9	7	6

The equation of the regression line calculated from the sample data can be used to convert each mark (x) awarded by the teacher into an equivalent mark (y) that the examiner would, on average, most probably have given to that same candidate. Such a moderation adjustment, based on extrapolating from a sample to a much larger collection of marks, can only reflect the general trends apparent in the marking. Individual variation relating to particular candidates cannot be accounted for. The purpose of moderation is to ensure that candidate marks, on the whole, are adjusted to more appropriate levels. Moderation cannot ensure a precisely correct outcome for every candidate.

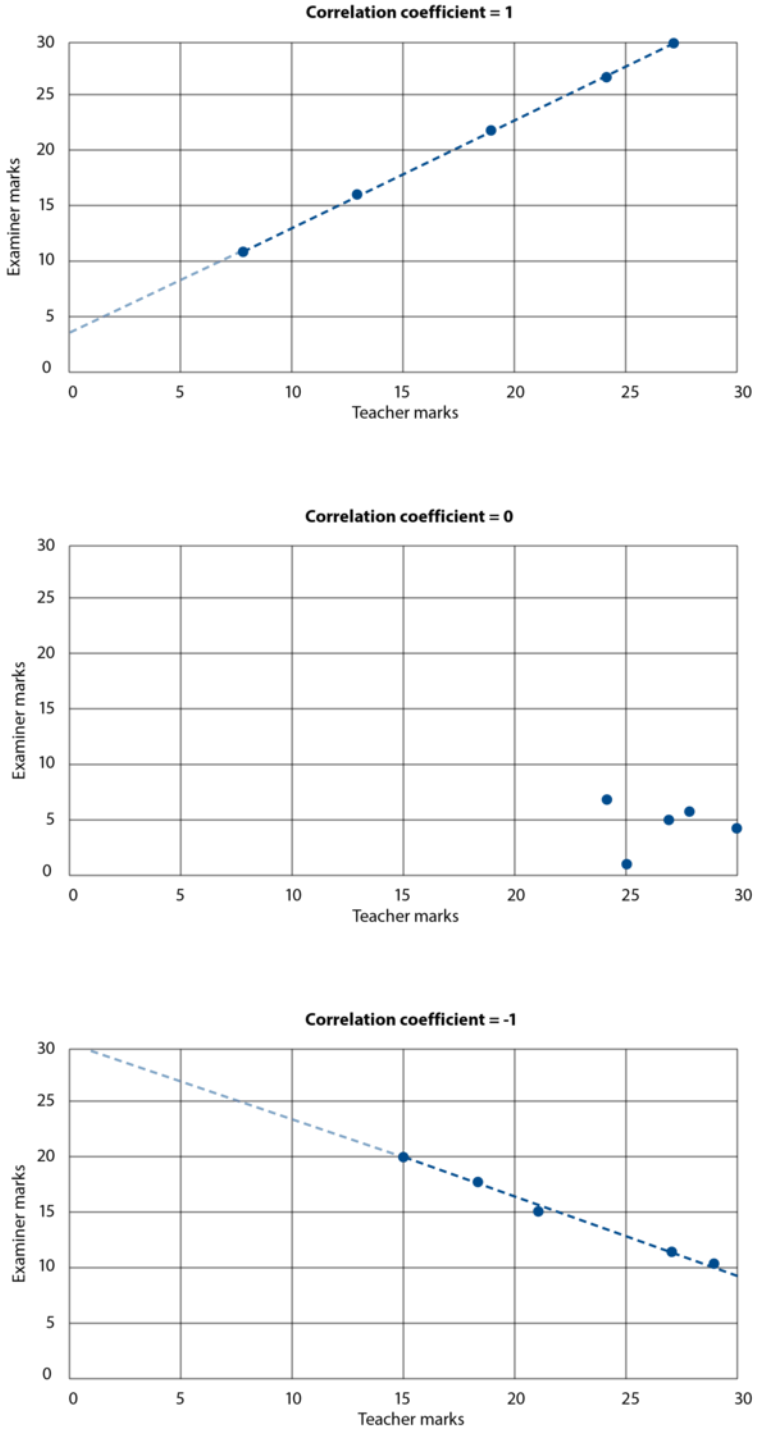
Moderation failure

A check is automatically carried out to make sure that the linear regression line of best fit (that is, the calculated moderation factor) meets certain conditions before it is applied to all of a teacher's marks. In some cases it may not be possible to calculate a moderation adjustment using the submitted sample work. One statistical measure is the correlation coefficient (the product moment correlation coefficient is used). This measures the consistency of the relationship between the teacher's and the examiner's marks.

- A correlation coefficient of zero indicates no relationship at all.
- A score of one indicates perfect consistency in the relationship between the marking and agreement in ranking candidates from best to worst (though not necessarily exactly the same marks).
- A coefficient of –1 indicates consistently opposing views with regard to the relative merits of candidates' work, with the teacher and the examiner producing opposite rankings to each other.

Figure AN3

Correlation coefficient graphs



For the calculated moderation factor to be acceptable, the correlation coefficient must be at least 0.85, indicating a high level of agreement between the teacher and the examiner. However, a high correlation coefficient on its own is insufficient to ensure the moderation factor is appropriate. A further check is carried out that the gradient (slope) of the regression line is between 0.5 and 1.5. If the gradient of the line is too low (or too shallow), it means that the teacher has spread candidates' marks out too much, giving

comparatively too few marks to weak work and too many marks to good work, even though this may be done on a consistent basis. The examiner has had to compress the teacher's mark range considerably. If the gradient is greater than 1.5, the line is too steep and the opposite applies; the teacher has not differentiated sufficiently between poor and good candidate work and the examiner has had to expand the mark range awarded.

A sample will "fail" the automatic moderation checks if the correlation coefficient is less than 0.85 or the gradient is outside of 0.5 to 1.5.

All cases of school samples which fail moderation are reviewed individually by IB assessment staff who consider the underlying data carefully and may decide:

1. that the calculated regression line is appropriate for the teacher's particular mark range
2. to apply some other moderation adjustment that is appropriate for the teacher's mark range
3. to request further sample data in order to clarify the trend
4. to request the rest of the candidates' work in order to carry out a complete re-mark of the teacher's marking.

All candidates' IA work must be available until the issue of results so we can resolve any moderation failures.

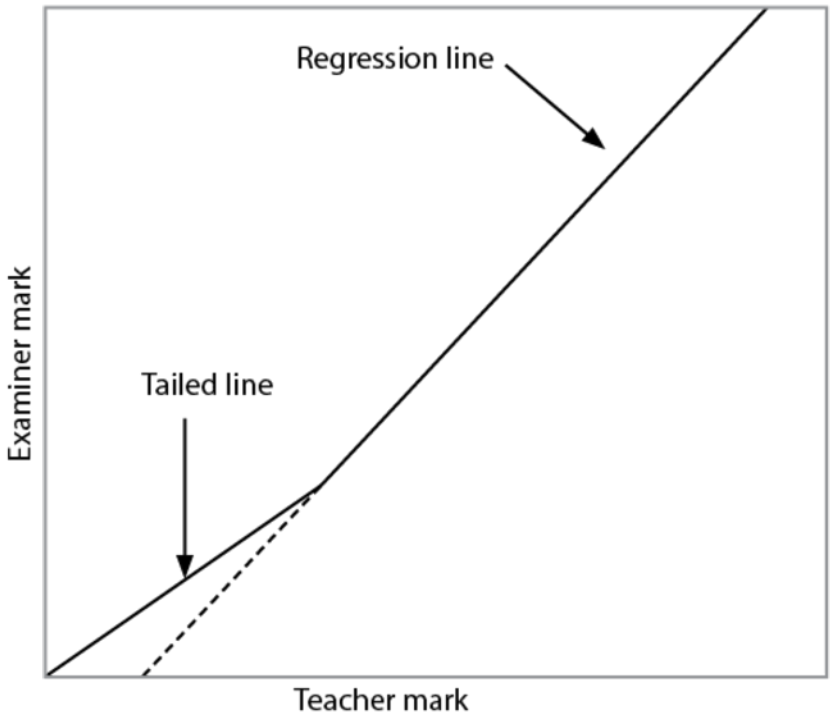
Adaption of linear model

The straight-line model used for moderation is modified to some extent by the use of "tailing". It can be seen that a straight line moderation adjustment may have inappropriate effects at the extremes of the possible mark range, making it impossible for any candidate to be awarded zero. A candidate whose work is genuinely worth zero might be given a few marks when nothing of any worth had been written on the work, or a candidate whose work is poor but worth a few marks may be given zero through moderation as would be the case from Figure AN3.

To overcome this problem, "tailing" is applied to marks in the bottom 20% of the available mark range. At this extreme, the calculated regression line is modified and substituted by a new "tailed" line that links from the regression line to the minimum coordinates, as shown in Figure AN4. Tailing of the regression line is not applied at the upper end because it frequently happens that teachers have awarded maximum marks to candidates who clearly do not deserve them.

Figure AN4

Example of tailed line regression



"Tailing" of a regression line to prevent candidate marks from being adjusted away from or towards minimum values.

Tailing ensures that a moderated zero mark can only be derived from an original zero mark. It prevents work that is worth a small number of marks from being given zero, and also prevents work that is worthless from being given a small number of marks. This assumes that the teacher's marking passes through the automatic moderation process. If the teacher's marking "fails" moderation and cannot be automatically moderated, tailing is not applied.

Annex 2: Roadmap for creating a validity argument

The IB needs to be confident that their courses and assessments are fit for purpose. This is not a simple exercise of a series of yes/no questions but a meaningful discussion of how competing priorities are balanced and the evidence that aims and intentions have been realized.

The "validity argument" is effectively the summary of these discussions, gathering together all the different strands to explain why the IB believes its courses and assessments are the best they can be. The topics and questions below form the starting point for demonstrating validity. Some answers relate to general processes in IB assessment such as qualification and seeding scripts, while others are specific to individual subjects and courses.

The process of creating a validity argument is not an end in itself, rather it is important that the right questions are asked, and evidence recorded, at each point in the cycle so that IB assessments and courses remain as good as they can be.

Developing the right course

- Is the curriculum appropriate and contemporary enough for students to understand the course?
- How is the course distinct from other IB courses?
- How do we know the curriculum can be taught effectively in the time allowed for teaching?
- The IB process of curriculum review gathers the evidence to answer these questions for the validity argument.

Aligning the objectives of the assessment with the aims of the course

What does it mean to be good at this subject? What does good look like?

What are we trying to assess with each task (component)?

How well do the aims of the course match what we are assessing?

How do the assessment criteria and grade descriptors reflect what good looks like for each task?

Examples of evidence would be explanation and justification of assessment mode, and notes of discussions around the criteria and descriptions.

Evaluating the assessment model

How do we know the curriculum and assessment standards (not performance standards) are comparable with other IB courses?

How does the assessment model minimize bias?

How might we adjust the assessments to engage with students with particular requirements, for example, blind candidates or those unable to attempt a particular task?

What is the acceptable level of professional disagreement when marking?

Evidence would be gathered during curriculum review and subsequent subject reports.

Reflecting on the process of creating examinations (specimen and live papers)

How do the questions accurately reflect the curriculum?

Examples of evidence would be the setting grid, which describes how the questions relate to the curriculum and assessment objectives, and scrutineer reports,

Special arrangements

Did the assessment arrangements cater for a wide range of candidates without requiring special arrangements?

Where special arrangements were required, were there any particular issues that had not been identified during the design phase?

How do we know that the special arrangement organized maintained comparability of the standard?

Examples of evidence would include decisions that lead to the creation of modified papers or evaluation of performance on modified questions compared to others that have not been modified.

Manageability of taking the assessments

Were there any problems with schools or with candidates that were a result of the IB's processes or the assessment design?

Evidence might include feedback from schools or candidates. No feedback may indicate that there are no issues.

Review of marking success

How reliable was the marking? Could the tolerances have been tighter?

Was the initial markscheme (produced with the examination) broadly correct or did it require significant amendments? Was this amendment reasonable or could the student responses have been foreseen at paper (and markscheme) authoring?

Examples of evidence include qualification and seed data and any replacement of definitively marked scripts.

Confidence and comparability in grading decisions

How confident was the CE in recommending grade boundaries?

Did all the evidence support the grade boundary decision or did different types of information suggest different conclusions?

How do we know that the grading was consistent with previous years and other subjects?

Evidence would include commentary in internal reports.

Evidence of fairness (bias)

How do we know that different groups of students were treated equally?

How confident are we that marking and grading was consistent between response languages and time zones?

Examples of evidence would include analysis of results data and teacher feedback. It would also include evidence from the Spanish and French examiners involved in the grade award, data on bilingual examiners on English response language seed scripts and review of candidate work from the other time zone in setting boundaries.

Impact of EURs

How confident are we that any EUR mark changes are reasonable?

How confident are we that the EURs demonstrate there was no systematic failure of marking or grading?

Examples of evidence would include analysis of any pattern in EUR changes and the evidence of an equal number of small mark changes up and down. The IB process of using seeding in EURs also supports this. Evidence of senior examiners engaging with these to support a consistent standard is also relevant.

Predictive validity

What evidence is there that the student's outcome is a good indication of future success in this subject? What evidence is there it is a good indication of success in related subjects?

Examples of evidence could include qualitative feedback from students and staff as well as quantitative evidence from the IB Research team.

Evidence of culture of continual improvement

What did we learn from the last session to improve this session?

What are the key lessons we can take from this session to inform future sessions?

How do we evaluate our success in making these improvements?

Evidence would include examples of how notes from previous sessions influence design this session.

Bibliography

Anderson, LW and Krathwohl, DR. 2001. *A Taxonomy for Learning, Teaching and Assessing: A Revision of Bloom's Taxonomy of Educational Objectives.* New York, USA. Longman.

Baird, J, Cresswell, MJ and Newton, P. 2000. "Would the real gold standard please step forward?" *Research Papers in Education*. Vol 15, number 2. Pp 213–229.

Black, P. 1999. "Assessment, Learning Theories and Testing Systems", in Murphy, P. 1999. *Learners, Learning & Assessment.* London, UK. Paul Chapman Publishing in association with The Open University.

Bloom, BS (Ed), Englehart, MD, Furst, EJ, Hill, WH and Krathwohl, DR. 1956. *Taxonomy of Educational Objectives, Handbook 1: Cognitive Domain.* New York, USA. Longman.

Broadfoot, P. 1996. *Education, Assessment and Society: A Sociological Analysis.* Buckingham, UK. Open University Press.

Brown, R. 2002. *Cultural dimensions of national and international educational assessment* in Hayden, M, Thompson, J and Walker, G (eds). *International education in practice: dimensions for national and international schools.* London, UK. Kogan Page.

Chamberlain, S. 2010. "AQA – Public Perceptions of Reliability". Ofqual's *Reliability Compendium*. Coventry. Office of Qualifications and Examinations Regulation. Chapter 18.

Coe, R, Searle, J, Barmby, P, Jone,s K and Higgins, S. 2008. "Relative difficulty of examinations in different subjects". *Report for SCORE (Science Community Supporting Education)*. Durham, UK. CEM Centre, Durham University.

Cresswell, MJ. 1986. "Examination Grade: How many should there be?" *British Educational Research Journal*. Vol 12, number 1.

Cresswell, MJ. 1996. *Defining, setting and maintaining standards in curriculum-embedded examinations: Judgemental and statistical approaches* in Goldstein, H and Lewis, T (eds). *Assessment: Problems, Developments and Statistical Issues.* Chichester, UK and New York, USA. Wiley. Pp 57–84.

Cresswell, MJ. 2000. "The role of public examinations in defining and monitoring standards". *Proceedings of the British Academy.* Vol 102. Pp 69–120.

Crooks, TJ, Kane, MT and Cohen, AS. 1996. "Threats to the Valid Use of Assessments". *Assessment in Education: Principles, Policy & Practice*. Vol 3, number 3.

Dolan, RP, Burling, K, Harms, M, Strain-Seymour, E, Way, W and Rose, D. 2013. "A Universal Design for Learning-based framework for designing accessible technology-enhanced assessments". (Research Report). Iowa City, IA. Pearson Education Measurement.

Frith, DS and Macintosh, HG. 1984. *A Teachers' Guide to Assessment.* Cheltenham, UK. Stanley Thomas.

Gibbs, G. 1992. *Assessing More Students*. Oxford, UK. Oxford Centre for Staff Learning and Development.

Gipps, C and Murphy, P. 1994. *A Fair Test? Assessment, Achievement and Equity.* Buckingham, UK. Open University Press.

Glaser, R. 1963. "Instructional technology and the measurement of learning outcomes: Some questions". *American Psychologist.* Vol 18. Pp 519–21.

Goldstein, H. 1996. *Group differences and bias in assessment* in Goldstein, H and Lewis, T (eds). *Assessment: Problems, Developments and Statistical Issues.* Chichester, UK and New York, USA. Wiley. Pp 85–93.

Good, FJ and Cresswell, MJ. 1988. *Grading the GCSE.* London, UK. Secondary Examinations Council.

He, Q, Opposs, D and Boyle, A. 2010. "A Quantitative Investigation into Public Perceptions of Reliability in Examination Results in England". Ofqual's *Reliability Compendium. Coventry*. Office of Qualifications and Examinations Regulation. Chapter 19. Page 68.

Hieronymous, AN and Hoover, HD. 1986. *Iowa tests of basic skills manual for school administrators*. Iowa City, USA. University of Iowa.

Hughes, D, Keeling, B and Tuck, B. 1983 "Effects of achievement expectations and handwriting quality on scoring essays". *Journal of Educational Measurement*. Vol 20, number 1. Pp 65–70.

Humphreys, LG. 1986. "An analysis and evaluation of test and item bias in the prediction context". *Journal of Applied Psychology*. Vol 71. Pp 327–33.

Lambert, D and Lines, D. 2000. *Understanding Assessment*. London, UK. Routledge Falmer.

Llewellyn, D. 2014. *Inquire Within: Implementing Inquiry and Argument-Based Science Standards in Grades 3–8* (third edition). Thousand Oaks, CA, USA. Corwin.

Linn, MC. 1992. "Gender differences in educational achievement". In J. Pfleiderer (ed). *Sex Equity in Educational Opportunity, Achievement, and Testing.* Princeton, NJ, USA. Educational Testing Service.

Meyer, A, Rose, DH, and Gordon, D. 2014. *Universal design for learning: Theory and Practice.* Wakefield, MA. CAST Professional Publishing.

Meighan, R. *Restructuring Education so it works for kids & society*. April 2014. Accessed 15 January 2017. http://lifelearningmagazine.com/0412/restructuring-education.htm

Mitra, S. 13–16 October 2011. *Speech to IB Heads World Conference.* Singapore.

Murphy, P. 1999. *Learners, Learning & Assessment.* London, UK. Paul Chapman Publishing in association with The Open University.

National Research Council. 2011. "Assessing 21st Century Skills: Summary of a Workshop. Committee on the Assessment of 21st Century Skills". Washington, DC. The National Academies Press.

Newton, PE. 2007. "Clarifying the purposes of educational assessment". *Assessment in Education: Principles, Policy & Practice*. Vol 14, number 2. Pp 149–170.

Newton, PE. 2012. "We need to talk about Validity". Paper presented to the National Council for Measurement in Education Annual Meeting. Vancouver, Canada.

Nuttall, DL, Backhouse, JK and Willmott, AS. 1974. "Comparability of standards between subjects: Schools Council Examinations Bulletin 29". *Schools Council*. Bulletin 29.

Ofqual. 2015. *Inter-Subject Comparability: A Review of the Technical Literature: ISC Working Paper 2*. Coventry, UK. The Office of Qualifications and Examinations Regulation.

Peterson, ADC. 1971. *New Techniques for the assessment of Pupils' Work*. Strasbourg, France. Council of Europe.

Peterson, ADC. 2003. *Schools Across Frontiers: The Story of the International Baccalaureate and the United World Colleges* (second edition). Chicago, USA. Open Court Publishing Company.

Popham, WJ. 1978. *Criterion-referenced Measurement.* Englewood Cliffs, NJ, USA. Prentice Hall.

RAND Education. *Teaching and Learning 21st Century Skills: Lessons from the Learning Sciences.* April 2012. New York: Asia Society. Accessed 1 January 2017. http://www.rand.org/pubs/external_publications/EP51105.html

Rao, K, Currie-Rubin, R and Logli, C. 2016. *UDL and Inclusive Practices in IB Schools Worldwide.* Wakefield, USA. CAST Professional Learning.

Schleicher, A. *The case for 21st-century learning*. 2016. Accessed 1 January 2017. http://www.oecd.org/general/thecasefor21st-centurylearning.htm

Shepard, LA. 1992. "Commentary: what policy makers who mandate tests should know about the new psychology of intellectual ability and learning", in Gifford, BR and O'Connor, MC (eds). *Changing Assessments: Alternative Views of Aptitude, Achievement and Instruction.* Boston, USA and Dordrecht, Netherlands. Kluwer. Pp 301–28.

Sireci, SG. 2007. "On test validity theory and test validation". *Educational Researcher*. Vol 36, number 8. Pp 477–481.

Smith, D and Tomlinson, S. 1989. *The School Effect*. London, UK. Policy Studies Institute.

Snyder, BR. 1971. *The Hidden Curriculum*. Cambridge, MA, USA. MIT Press.

Surgenor, P. 2010. *Teaching Toolkit: Effect of Assessment on Learning*. Dublin, Ireland. UCD Dublin.

Vygotsky, LS. 1962. *Thought and language*. New York, USA. Wiley.

Wiliam, D. 1993. "Validity, dependability and reliability in National Curriculum assessment". *The Curriculum Journal.* Vol 4, issue 3. Pp 335–350.

Winkley, J and Cresswell, M. 2010." Introduction to the concept of reliability", Ofqual's *Reliability Compendium.* Coventry. Office of Qualifications and Examinations Regulation. Chapter 1.

Wood, D, Bruner, JS and Ross, G. 1976. "The role of tutoring in problem solving". *Journal of Child Psychology and Psychiatry*. Vol 17. Pp 89–100.

Wood, R. 1991. *Assessment and Testing: A survey of research.* Cambridge, UK. Cambridge University Press.

Wood, D. 1998. *How Children Think and Learn: The Social Contexts of Cognitive Development* (second edition). Oxford, UK. Blackwell.

OECD. *21st century skills and competences in OECD countries*. December 2009. Accessed 31 January 2017. https://www.oecd.org/edu/ceri/44254873.ppt

Glossary

Term	Definition
Academic honesty	A set of values and skills that promote personal integrity and good practice in teaching, learning and assessment.
Academic misconduct	Behaviour (whether deliberate or inadvertent) that results in, or may result in, the candidate or any other candidate gaining an unfair advantage in one or more components of assessment. Behaviour that may disadvantage another candidate is also regarded as academic misconduct.
Achievement level	The level given when the student work reflects the corresponding descriptor. Achievement levels are shown in the left-hand column of the assessment criteria.
Adverse circumstances	Adverse circumstances are defined as those beyond the control of the candidate that might be detrimental to his or her assessment performance, including severe stress, exceptionally difficult family circumstances, bereavement or events that may threaten the health or safety of candidates. The same circumstances may affect a group of candidates or all candidates within a school. Adverse circumstances do not include: • shortcomings on the part of the school at which the candidate is registered. • the failure of candidates to improve performance despite receiving authorized inclusive access arrangements.
Aggregation	The process of combining marks and scores into a final outcome.
Alignment	Agreement in principle and practice between shared values and aspirations for learning (written curriculum), how teachers actually work (taught curriculum) and what students actually learn (assessed curriculum).
Analytic markscheme	A markscheme which tells you what the right answer is and where marks should be awarded.
Assessment	The collection of evidence in order to make judgments about teaching and learning.
Assessment access requirements	A candidate with assessment access requirements is one who requires changes in assessment conditions to demonstrate his or her level of attainment.
Assessment component	An assessment component is made up of one or more tasks that are collected together, comprising part of the overall assessment. For example, an examination paper, portfolio of work, project or research assignment.
Assessment criteria	Criteria against which a student's performance is measured.

Term	Definition
Assessment cycle	The steps taken in creating, taking and marking assessments including examinations. It is a cycle because the IB learns from each examination session to improve future sessions.
Assessment response	A term used to describe all material produced by a candidate in response to assessment material.
Assessment strategy	The method or approach that is used when gathering information about student learning (for example, observation, open-ended tasks, selected responses).
Assessment task	The activity or series of activities with which students engage in order for assessment to take place.
Assessment tool	A method of collecting information about a learner's performance and understanding.
Atypical response	An answer to a task which is significantly different to those usually received. Examples of atypical responses include incomplete work, noncompliant work, unanticipated responses, problematic work or malpractice.
Authentication	Process and proof that the work has been undertaken by the candidate. Examples include signatures from the teacher and candidate that provide provenance for the candidate's response.
Backwash effect	The impact which later parts of a process have on the delivery of earlier parts. In the educational context backwash usually refers to the way teaching and learning is changed by how the candidate is assessed.
Bias	Bias is where a defined group (that is, racial or ethnic group or gender) performs differently on a specific question or task than the average for a reason other than ability in the trait that is being assessed.
Candidate	A student registered for assessment.
Candidate registration	Process undertaken by school coordinator to register candidates for IB assessment.
Chief examiner	The most senior examiner who is responsible for ensuring that standards are maintained over time and between disciplines within a subject group (for example, sciences).
Command term	The word(s) in a question which explain the assessment objective which is being assessed.
Comparability	The degree to which a particular outcome can be considered the same as another outcome. It is typically used between years (is a grade 8 this year the same as a grade 8 last year?) or between subjects (is a grade 5 in maths the same as a grade 5 in art?).
Component	See **Assessment component**.
Construct relevance	The degree to which the assessment actually tests the skills and knowledge that it is intended to. An example of a low level of construct relevance would be testing a student's practical skills with a written exam.
Course	A prescribed number of classes, lessons or teaching hours within a defined period of study. Schools organize teaching and learning of subjects through disciplinary and interdisciplinary courses.

Term	Definition
Course results	Course results is the primary outcome document. It shows each discipline the candidate has taken and the grade achieved (1–7). It also shows the grade achieved in the core components, interdisciplinary assessment and that the school's community service requirement has been completed. Finally the document records candidate's name, personal code, session number, session in which the awards were achieved, date of issue, name of school registering the candidate (and replacement, if appropriate). The results document only shows positive achievement.
Criterion-related assessment	An assessment process based on determining levels of achievement against previously agreed criteria. The standard is therefore fixed rather than depending on the achievement of the entire cohort of students.
Criterion-referencing	A comparison of student attainment against pre-defined descriptions of achievement (criteria) for grading.
Definitive mark	The mark awarded by the principal examiner for a particular piece of student work. This represents the mark that every other examiner should be aiming to replicate. (See also the **Quality model**)
Achievement level descriptors	Achievement level descriptors describe the features of student work expected to be seen at each achievement level.
Differentiation (in assessment)	To distinguish between candidates demonstrating different levels of competency.
Differentiation (teaching and learning)	Modifying teaching strategies to meet the needs of diverse learners, through varied content, process and products.
Discipline	A branch of learning or field of academic study; a way of ordering knowledge for the purpose of instruction (known generally for practical purposes of assessment in the MYP and DP as subjects). Some MYP subject groups and subjects can comprise multiple disciplines. For example, the MYP subject group arts includes disciplines like visual arts, drama, music, media and dance. The subject integrated sciences includes three disciplines: biology, chemistry and physics.
Dynamic sampling	A refinement of the moderation process, which allows better use of quality checks. It applies the "tolerance" quality model to both teachers' and examiners' marks. For teachers, if the initial sample is within tolerance, then no moderation factor will be applied. It also means moderators (examiners) receive student work individually which allows for "seed scripts" to be included to maintain a consistent standard. It also permits examiners to be allocated the necessary additional scripts if there is evidence that the teacher marking does not match the overall standard.
eAssessment	Assessment carried out on a computer or similar device.
eMarking	The process by which examiners mark examination material directly on the computer screen
Enquiry upon results (EUR)	Review of level (marks) undertaken at a school's request.
ePortfolios	The system/process by which schools upload candidates' internally assessed examination/coursework material to be externally moderated by the IB.
eScript	The candidate's responses (answers) to an eAssessment.

Term	Definition
Examination	A collection of one or more tasks of various types (short answer, extended answer, problem-solving or analytical questions; sometimes practical or oral tasks) that students must respond to under controlled conditions in a set time.
Examination invigilator	Individual who supervises and controls the exam environment.
Examination paper	The set of tasks and questions which a candidate is asked to complete. In certain circumstances it may refer to an examination which is taken on-screen, or an examination taken with pen and paper.
Examination session	The period during which exams are taken and marked. The IB has two examination sessions a year in May and November.
Examiner	Individual who assigns marks to the candidate's external assessment.
Examiner re-mark	The process of re-marking an examiner's allocation of eResponses where their marking is found to be inconsistent or deviates significantly from the required standard. This often occurs as a result of moderation failure.
External assessment	Assessment that is set and marked by the IB and not by a student's teacher.
Exceptional circumstances	Circumstances that are not commonly within the experience of other candidates with assessment access requirements. The IB reserves the right to determine which circumstances qualify as "exceptional" and therefore justify a particular inclusive access arrangement.
External moderation	See **moderation**.
Externally assessed	Work that is assessed/marked wholly by the IB.
Familiarization tool	A generic simulation of an examination that candidates can take in order to learn how to use the on-screen examination environment and toolsets.
Final assessment	The summative assessment of student work at the end of the period of study.
Formative assessment	Ongoing assessment aimed at providing information to guide teaching and improve student performance.
Grade	The description of student achievement. Final grades for student work range from 1 (lowest) to 7 (highest). The grade represents the IB's judgment on the overall qualities that the candidate has demonstrated and is consistent between years and subjects.
Grade award	The grade award process decides how to convert marks into grades to ensure that grades should mean the same thing whichever session a student takes their exam in.
Grade boundary	The point at which candidate achievement moves from one grade to another. It is often used to indicate the lowest or highest criterion level totals which corresponds to a particular grade.
Grade descriptors	The articulation of the qualities expected of students to achieve each grade. A grade descriptor may be specific to a subject, specific to a subject group, or general across a whole programme. In each case, a grade descriptor should describe the same characteristics; the more specific examples only explain what these descriptions mean in a subject-specific context.

Term	Definition
Holistic criteria	Approach to evaluating a candidate's work which considers the work as a single outcome, rather than looking at separate elements of it individually (for example, communication, subject knowledge, quality of argument and so on).
IB Information System (IBIS)	A system that allows school coordinators to complete administrative procedures and obtain news and information from the IB via a password-protected web server.
Inclusive access	An assessment that has considered the needs of all candidates, so that candidates can fairly demonstrate their competence in the subject.
Inclusive access arrangements	Changed or additional conditions during the assessment process for a candidate with assessment access requirements. These enable the candidate to demonstrate his or her level of attainment more fairly and are not intended to compensate for any lack of ability.
Interdisciplinary assessment	Combining or involving two or more branches of learning or fields of academic study within a single assessment. In the DP an interdisciplinary subject is one that meets the requirements of two subject groups through a single subject. In the MYP, interdisciplinary study can be developed both within and between/among subject groups. MYP external interdisciplinary assessment always involves multiple subject groups.
Internal assessment	Assessment carried out by teachers in the school.
Internal standardization	The process by which all teachers of a particular subject in a school ensure they are assessing to the same standard.
Internally assessed	Work that is assessed (marked) by the students' teachers. Internally assessed material is sampled by the IB for moderation purposes.
Issue of results	The process of candidates receiving grades from IB based on their assessments.
Item	Smallest unit of an assessment task or question. Each item generates a number as the mark. An item could be a whole question or parts of a question.
Judgment	The consideration of a candidate's work against an individual assessment criterion.
Maladministration	Maladministration is an action by an IB World School that infringes IB rules and regulations and potentially threatens the integrity of IB exams and assessments. It can happen before, during or after the completion of the assessment or completion of the examination.
Malpractice	Any practice which subverts the principles of academic honesty (for example, plagiarism).
Manageability	The degree to which the assessment and individual tasks place a burden on the student or school. Examples of manageability include the length of the assessments, the equipment or material required to deliver the assessment or the number of assessments require in a qualification.
Mark(s)	Value that reflects the quality of the candidate's answer to the specific question asked.
Markbands	A specified/specific range of marks that should be awarded to a candidate answer that shows certain qualities.

Term	Definition
Markscheme	Guidance for awarding criterion levels for a given piece of work.
MCQ	See **Multiple-choice question**.
Missing grade procedure	A mechanism for providing a grade for students where the IB is not able to access an accurate or fair grade based on the work the candidate has completed. It is appropriate in those circumstances where the reason for the lack of evidence is due to the actions of the IB or third parties (not including the school) where it would not be reasonable for the student to be asked to complete the assessment on another occasion.
Missing mark procedure	A mechanism for providing a mark for students where the IB is not able to access an accurate or fair mark based on the work the candidate has completed. It is appropriate in those circumstances where the reason for the lack of evidence is due to the actions of the IB or third parties (not including the school) where it would not be reasonable for the student to be asked to complete the assessment on another occasion.
Moderation	A process to ensure that a common standard of assessment is achieved through review of samples of assessed student work and adjusting assessors' assessments where necessary.
Moderation factor	An arithmetical adjustment applied to an assessor's criterion levels total to bring them in line with the common assessment standard.
Moderation sample	The sample of student work submitted to the IB to ensure it is marked to the required standard.
Modified paper	Changes made to an assessment to allow a student with specific needs to be able to take the assessment on an equal footing with students who do not have these needs. Examples include changing the shape or style of the type font. Such adjustments must not change the nature of the question being asked.
Multiple-choice question	A question where a candidate must select the correct answer from a list of supplied possible answers.
Norm-referencing	Where attainment is determined by comparing (referencing) to the candidate's performance against that of the entire population for whom the assessment is designed.
Objective	One of a set of statements describing the skills, knowledge and understanding that will be assessed.
On-screen examination	A formal, timed, externally produced, media-rich examination comprising a series of tasks related to the subject designed to be answered in a secure exam environment.
Paper	See **Examination paper**.
Paper author	A person who creates the questions and associated markscheme that will be used for the assessments.
Pilot subject	A subject undergoing evaluation, which pending successful evaluation will become generally available.
Plagiarism	The representation, intentionally or unintentionally, of the ideas, words or work of another person without proper, clear and explicit acknowledgment.

Term	Definition
Practice script	Examples of student work that are identified and marked during standardization and then given to examiners to explain that this is the standard that they should be marking to.
Predictability	Predictability is the state of being able to gauge what and/or when something will happen. In the context of assessment, this means the ability of schools to anticipate questions that will be asked on a paper, and when. Good predictability is essential for IB working practices in assessment as, by adhering to it, it means the IB remains loyal to the requirements of their constructs, as published for teachers, leading to a "fair" assessment opportunity in terms of curriculum alignment (for example, what the IB said would be assessed is assessed).
Principal examiner	The principal examiner (PE) is responsible for leading the assessment of a component. They set the standard for the assessment and are usually also one of the assessment authors. In the MYP, the role of principal examiner is slightly different from other examination systems. A principal examiner is the head of a particular discipline and is responsible for leading the team designing the assessment, for setting and maintaining standards and mentoring examiner team leaders.
Qualification script	An example of student work selected by the principal examiner used to formally check that examiners have understood the required standard of marking before they are allowed to mark live student scripts.
Quality model	The approach that the IB takes to ensure that students receive the correct assessment outcome. The principal examiner sets the correct standard of response for each question and each examiner needs to reproduce this standard. For externally marked assessments, this is done by providing guidance to examiners though standardization, checking their understanding of the standard with qualification scripts and then monitoring their marking regularly through seed scripts.
Question	Task or activity used to allow a candidate to demonstrate their competence in a subject.
Question bank	Collection of questions provided with information about the topic and expected degree of difficulty. The information in a question bank can be used to create examination papers. The IB does not currently use question banks.
Question item group (QIG)	One or more related questions within an examination paper are considered as a group. Examiners are then asked to mark individual QIGs rather than whole papers. This approach provides more reliable marking than whole script approaches.
Reliability	The degree to which the candidate will receive the same outcome every time his or her work is assessed. It can refer to the reliability between examiners (that is, do they give the same outcome for the student?) or the reliability of a single examiner (that is, do they give the same outcome every time he or she looks at the student's work?).
Response language	The language in which the student answers the assessment.

Term	Definition
Retake	A second or subsequent attempt at one or more examinations in the hope of obtaining an MYP or DP certificate or increasing the total mark on a certificate already received.
Script	The candidate's answers to the exam paper. It can also be any candidate work which has been submitted for assessment.
Seed	A seed is a script that has already been marked by the principal examiner and is randomly added to a batch of scripts allocated to an examiner for marking. It looks like any other script so the examiner cannot tell it is a seed. The marks the examiner awards the seed will be checked against those given by the principal examiner, with a certain tolerance to check the examiner is marking to the set standard. Dynamic sampling moderation seeds are used in the same way as part of the moderation process.
Senior examiner	A role describing experienced examiners who support the principal examiner.
Session	See **Examination session**.
Special consideration	A candidate affected by adverse circumstances may be eligible for special consideration, provided that this would not give an advantage in comparison with other candidates. In such cases if the candidate is within one or two scaled marks of the next higher grade boundary, the candidate's grade in the affected discipline(s) will be raised. This is the only possible accommodation for candidates in the event of adverse circumstances. If a candidate's marks are not within the required range, then no adjustment will be made. When a candidate is affected by adverse circumstances, he or she may be eligible for special consideration, provided that this would not give an advantage in comparison with other candidates. In such cases, if the candidate is within one or two scaled marks of the next higher grade boundary, the candidate's grade in the affected discipline(s) will be raised. This is the only possible accommodation for candidates in the event of adverse circumstances. If a candidate's marks are not within the required range, then no adjustment will be made.
Standard	The performance which is expected to achieve a particular score, grade or outcome.
Standardization meeting	A meeting held by the principal examiners to describe the required standard for marking and set seed scripts.
Standardization	The collaborative process by which a common standard of assessment is achieved among moderators or examiners.
Student	A person who is taking part in an IB course or educational programme. Students become candidates when they are part of the assessment process.
Submission	The candidate (or school on behalf of the candidate) handing in their final work to the IB.
Summative assessment	Assessment aimed at determining the competency or level of achievement of a student generally at the end of a course of study or a unit of work.

Term	Definition
Teacher support material	Additional information to help teachers understand what is required by the IB course. It is intended to give practical help to aid understanding and implementation of the theory in the subject guides.
Team leader	An examiner who leads a team of examiners.
Tolerance	The small variation from the principal examiner's definitive mark, which the IB believes is close enough to show the examiner is still marking to the correct standard. Tolerances are necessary because marking is a matter of judgment and even experienced markers will vary slightly when re-marking the same piece of student work. Tolerances vary according to the number of marks, the kind of question and the subject.
Universal Design of Assessment	The concept of Universal Design of Assessment is that all assessments should be developed with an understanding of the range of requirements that candidates may have, rather than to treat some candidates differently. This is part of IB's commitment to Universal Design for Learning (UDL).
Validity	The overall term that describes whether an assessment or the purpose for which the assessment results are being used is fit for purpose.
Validity argument	The evidence and explanation for decisions made in creating an assessment which justifies it is fit for purpose.
Weak criterion-referencing	If student attainment is compared against pre-defined descriptions of achievement (criteria) and the performance of previous cohorts, this is known as weak criterion-referencing (or criterion-related referencing). This is the approach the IB takes to maintaining standards.
Working languages	The languages in which the IB communicates with its stakeholders and in which it is committed to providing a range of services for the implementation of the programmes. They are currently English, French and Spanish.