Diploma Programme assessment

Principles and practice

Diploma Programme assessment

Principles and practice

International Baccalaureate Organization

Buenos Aires **Cardiff** **Geneva** **New York** **Singapore**

Diploma Programme assessment
Principles and practice

International Baccalaureate Organization, Geneva, CH-1218, Switzerland

First published in September 2004
Updated November 2010
by the International Baccalaureate Organization
Peterson House, Malthouse Avenue, Cardiff Gate
Cardiff, Wales GB CF23 8GL
UNITED KINGDOM

Tel: + 44 29 2054 7777
Fax: + 44 29 2054 7778
Web site: www.ibo.org

IBO merchandise and publications in its official and working languages can be purchased through the online catalogue at www.ibo.org, found by selecting *Publications* from the shortcuts box. General ordering queries should be directed to the sales department in Cardiff.

Tel: +44 29 2054 7746
Fax: +44 29 2054 7779
E-mail: sales@ibo.org

Printed in the United Kingdom by Antony Rowe Ltd, Chippenham, Wiltshire.

DR28

The IBO mission statement

The International Baccalaureate Organization 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 IBO 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.

Contents

1. Introduction and overview

This booklet is intended to explain the workings of the International Baccalaureate Diploma Programme assessment system and their underlying rationale. The Diploma Programme (DP) is taught in over 100 countries by schools representing a wide variety of educational contexts and traditions. To some of these schools, the philosophy and approaches adopted by the International Baccalaureate Organization (IBO) in assessing their students will seem familiar, while to others the system might seem mysterious and obscure. It is important that teachers who prepare students for entry to DP examinations should understand the nature of the assessment system to which they are submitting their students, as well as the educational philosophy of the programme as a whole. Nearly all DP teachers necessarily participate in the assessment process, making a direct contribution to the final assessment carried out on their own students. In addition, many DP teachers have further involvement as examiners in marking other students' work, in checking the marking of other examiners, or even by contributing to the writing of examination papers (only for papers that will not be taken by their own students). For all of these reasons, an appreciation of how their own contribution fits into the general scheme of assessment can only be helpful for DP teachers.

In addition to its role as a support document for DP teachers and coordinators, this publication may also prove useful for university admissions officers, school heads, school governing bodies, examiners and students themselves. While there are many universities around the world that are now very familiar with DP applicants and the qualities they bring with them (see, for example, IBO, 2003a), there still remain a great many more to whom the IB diploma qualification is unknown. Many students and their parents/carers will also have a keen and legitimate interest in understanding the nature of the assessment system that is such an integral part of participation in the IB Diploma Programme.

This booklet will focus on the formal aspects of assessment within the DP: those that contribute to the final qualification. It is not the purpose of this booklet to discuss in detail the equally important area of formative assessment, by means of which classroom teachers can have a more direct and immediate impact on the progress of their students' learning. Teachers wishing to discover more about formative assessment are advised to consult works such as Black and Wiliam (1998a), Black and Wiliam (1998b), Assessment Reform Group (1999), and Sadler (1998), although the topic is also dealt with in most standard books on educational assessment.

In order to understand the nature of the DP assessment system, it is necessary to provide some background on the historical and theoretical development of assessment practice. The approach adopted in this booklet is necessarily one of survey and brief description. Many significant issues are presented in simplified form. However, it is important that historical and conceptual issues should be raised, because they have a significant impact on current practice. For readers who wish to find out more about the issues and challenges that face those with responsibility for constructing and administering assessment systems, the works quoted in the text will make a suitable starting point for further investigation. Further details of the actual mechanisms and procedures of DP assessment exist in the many internal procedure documents produced by the IBO, but only those directly involved in conducting assessment are likely to need this level of detail.

The next section of this booklet, *Principles of assessment*, reviews the theoretical concepts that have guided the development of large-scale educational assessment over recent decades. Some commonly used assessment terminology is introduced and explained. The different backgrounds of the two main approaches to assessment—psychometric testing and performance assessment—are described, together with the varied purposes for which assessment results are used. This section closes by emphasizing the compromises that have to be made between conflicting requirements when designing an assessment system, as there are strong tensions between reliability, aspects of validity, equity and manageability.

Section 3, *Diploma Programme assessment—aims and approaches*, shows how the concepts and principles described in the previous section are taken into account in constructing the approaches to assessment used in the DP. Although the reliability of the final results, and the validity of the assessments in terms of the way they address appropriate cognitive skills through a wide range of different assessment tasks, are important, neither of these concerns overrides the primary aim of devising an assessment system that supports and encourages appropriate teaching and learning in the classroom. The other main principles of DP assessment are that results should be based on the professional judgment of senior examiners, and assessments should reflect the international/intercultural dimensions of the programme.

In section 4, *Diploma Programme assessment structures*, the main assessment structures are briefly outlined. These structures are:

- the curriculum model, describing the range of courses that are assessed and that complement each other to make up a balanced and integrated programme of study for each student following the full Diploma Programme
- the different assessment models applied to each course of study, in which internal assessment carried out by the classroom teacher plays a significant role
- the different groups of people—IBO staff and associated teachers, examiners and other academics, who have different roles to play in devising and conducting the assessments.

The final section provides a broadly chronological account of *Diploma Programme assessment processes*, from the preparation of the examination questions for each session, through the marking process and the awarding of grades, to the release of final results to students and the follow-up that occurs afterwards. The total period covered by all these activities is about two years, and since there are two sessions of examinations each year, in May and November, there is considerable overlap in the activities relating to each session.

This booklet is intended as a reference document for the different types of reader mentioned at the beginning, who are all stakeholders in assessment for the DP and who may wish to consult different parts of the booklet as issues or questions arise. It is hoped that, as well as explaining how DP assessment works, the booklet will also provide an understanding of why assessment is conducted in such a manner. Those readers whose main concern is how DP assessment works will find sections 4 and 5 of most interest, while readers who are interested in the underlying theoretical principles will hopefully find sections 2 and 3 informative.

Further explanations of the different aspects of validity and reliability, together with the DP assessment policy that governs the development of assessment models for each subject, are provided in the appendices. Readers who have queries beyond the scope of this booklet are invited to contact IBO assessment staff, by e-mailing assessment@ibo.org.

2. Principles of assessment

2.1 Why "assessment"?

For many people the words "assessment", "examination" and "test" have a similar meaning and are used somewhat interchangeably. For the purposes of this booklet, more specific meanings are necessary and the following will be adopted throughout.

Test—a collection of many short-answer questions (either selected-response/multiple-choice questions or questions requiring only a few words in response) that students must answer under controlled, isolated conditions in a set time. Often marked (or graded) automatically.

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, isolated conditions in a set time. Generally marked/graded by examiner (or rater).

Assessment—a term used to cover all the various methods by which student achievement can be evaluated. Assessment instruments may include tests, examinations, extended practical work, projects, portfolios and oral work, some carried out over a prolonged period and sometimes marked by the student's teacher.

A distinction is often made between *summative* assessment, aimed at determining the level of achievement of a student generally at the end of a course of study, and *formative* assessment, aimed at identifying the learning needs of students and forming part of the learning process itself. Although these two functions are apparently quite distinct, the same assessment instruments can often be used for either purpose, the difference lying in the way the outcomes of the assessment are interpreted and applied (Black, 1993a; Wiliam and Black, 1996). Biggs (1998) has also made it clear that it is not helpful to regard formative and summative assessment as being mutually exclusive. The two approaches should interact and be mutually supportive. In the context of the Diploma Programme (DP), the term *formal assessment* is preferred to describe all those assessment instruments that are used to contribute to the final qualification. Some of these instruments can be used formatively during the course of study as well as summatively towards the end of it, an approach that has been proposed elsewhere (for example, Lambert and Lines, 2000, Ch 10).

Formal assessment of the DP includes some multiple-choice tests for a few subjects and examination papers for most subjects, intended to be taken at the end of the two-year course, and a variety of other tasks (essays, research essays, written assignments, oral interviews, scientific and mathematical investigations, fieldwork projects and artistic performances) spread over different subjects and completed by students at various times under various conditions during their course. The reason for the variety of forms of assessment instrument will be explained in section 3.5.

2.2 Formative and summative—what are the purposes of assessment?

Assessment can be used for a variety of purposes. The intended purpose for a given system of assessment will have a major impact on its style and format. For formative assessment, the main purpose is to provide detailed feedback to teachers and their students on the nature of students' strengths and weaknesses, and to help develop students' capabilities. Methods of assessment involving direct interaction between teacher and student are particularly helpful here. The teacher is seen as a supporter rather than a director of learning (Vygotsky, 1962; Vygotsky, 1978), and should make use of assessment tasks and instruments that help the student work in what Vygotsky refers to as the "zone of proximal development". This is the range of achievement between what the student can do on his/her own, and what the student can do with the support of the teacher. This is a similar concept to the notion of "scaffolding" 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.

It is more important that formative assessment correctly identifies the knowledge, skills and understanding that students should develop, rather than accurately measuring the level of each student's achievement. *Reliability* is therefore a much lower consideration for formative assessment than *validity* (see section 2.6 for explanations of the terms "reliability" and "validity").

Summative assessment is used for quite different purposes, including the provision of information about student achievement, the certification and selection of students, an accountability mechanism to evaluate teachers and schools, and a driving force for reform of curriculum. The use of summative assessment as an accountability mechanism, with its associated goals of raising standards and providing information on which to identify "good" schools and teachers, is a controversial issue (Gipps and Stobart, 1993; Goldstein, 1996b). The controversy surrounds, on the one hand, the difficulty of making fair comparisons between teachers and schools that may have students from very different backgrounds and may be teaching in very different contexts, and on the other hand, the difficulty of interpreting apparent rises in standards of performance. Such rises may reflect genuine improvements to teaching and learning, or more concentrated efforts at teaching to the test, with an inferred corollary of increasing neglect of other aspects of education.

There is no formal role for DP assessment as an accountability mechanism by which school performance is judged, although schools themselves and parents of students may use diploma results in this way. The IBO considers it to be largely the individual school's role to evaluate its own effectiveness. A five-year programme evaluation of authorized DP schools is conducted, primarily through self-study. Only rarely is student performance on DP assessment a major consideration in evaluating the authorized status of a school, although recurrent maladministration of assessment or improper conduct on the part of the school may lead to a review of that school's authorization to teach the programme (IBO, 2003b). The policies applied by DP schools in the way they decide which students should enter for the diploma, and the social and educational contexts in which the schools operate, are so varied that it would be inappropriate to judge school effectiveness solely on the basis of DP examination results.

Although the primary role of DP assessment is generally perceived to be that of certification of achievement, leading in most cases to a selection process for university admission, the other uses of summative assessment are also of significance. As far as is practicable, DP assessment is also seen as a major tool for reinforcing the teaching of the curricular goals of the programme—indeed, such assessment can only be valid if it adequately reflects these goals. A third purpose, that of providing differentiated information about student achievement (and hence teacher effectiveness) to inform the professional development of teachers, is also of significance in contributing to the impact of the DP on the education of students. The fulfillment of these different roles, which make differing demands on an assessment system, inevitably leads to some compromises in the design of DP assessment, as discussed in section 4.

It is worth noting at the outset 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 associated 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 (Cresswell, 1996; Broadfoot, 1996).

2.3 Psychometric testing and performance assessment

There are two major, and very different, traditions in assessment that are currently in wide use in national education systems. The first of these, psychometrics, has its roots in the development of intelligence testing in Paris at the beginning of the 20th century (Binet and Simon, 1905). The intent of psychometric testing is to use a number of carefully calibrated short questions (generally multiple-choice questions) to accurately measure a student's aptitude or potential in a particular area, for example, reading or arithmetic. There is an underlying assumption in this approach that the aptitude being measured is a fixed and unchanging attribute of a person. Based on this assumption, accuracy of measurement in terms of a precise ranking of students is considered the most crucial factor, to the extent that only those items that discriminate most effectively between different students are retained for use in the tests; other items, that may not discriminate between students so well, are discarded, regardless of the educational value of question content. Every question on the test is required to assess the same skill or trait, and so relative student performance on each question is assumed to reflect a linear measure of ability. Thus, strong statistical relationships are expected between performance on each question and performance on the whole test. Such tests are generally referred to as standardized objective tests and are taken under strictly controlled conditions.

In addition to measurement accuracy, equity has also been a major consideration in the development of these tests. Because the tests are automatically marked there can be no unfairness introduced by different standards of marking/grading. The main intent is to measure student capability, regardless of social or educational background, or ethnicity. Strenuous efforts are made to identify and exclude any items that are biased towards or against a particular sub-group of students. Difficulties surrounding this issue are discussed in section 2.7.

The primary intention to measure aptitude gradually shifted with increasing realization that in educational tests it was impossible, and often not desirable, to separate latent ability from the effects of educational experience. There was also a growing appreciation that latent abilities, such as intelligence, are not fixed unidimensional properties of an individual. Measuring a person's intelligence is not the same as finding their height, for example, which is a process that may justify continuing refinements to the accuracy of measurement.

Standardized tests came to be used more and more to measure achievement. However, the highly statistical approach remained, as did the style of questions. Student achievement against the desired curricular goals of classroom teaching came to be inferred from their performance on a very restricted kind of assessment task. Such assessment of broader achievement (for example, writing skill) by inference from a restricted kind of achievement (multiple-choice question responses) can only be effective as long as the assessment is "low-stakes" and students and teachers do not give much attention to the difference between what should be taught and what is being tested.

However, psychometric testing has become a very high-stakes operation in some countries, linked to school and to teacher accountability, with the outcome having a significant impact on students' life chances. In such an environment, the limited nature of skills and knowledge tested will inevitably have a major impact on the way teachers teach in the classroom, the so-called "backwash" effect. The negative effect of such teaching to the test has been identified as a significant disadvantage of standardized objective testing (for example, Hoffmann, 1962; Resnick and Resnick, 1992).

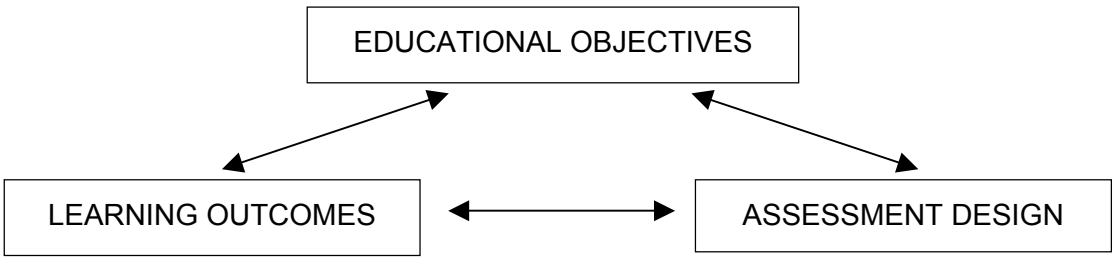
The belief that tests can and should measure a single, unified attribute has been open to question—even the concept of intelligence is now widely regarded as multidimensional (see, for example, Gardner, 1983). Apart from the narrow uniformity of approach, it is seen as particularly unfortunate that standardized tests encourage an atomized and disconnected view of learning. However, standardized testing does have a number of significant advantages, such as cost effectiveness and time efficiency when used on a large scale, and also a high degree of reliability of outcome.

The other main tradition is performance assessment, a term that refers to a wide variety of forms of assessment in which students are expected to carry out tasks that directly reflect the range of knowledge and skills they have learned in the classroom. Thus, performance assessment includes problem solving, essays, project work and examinations.[1] Because of the need for variety of assessment task and the construction of a substantial response (or performance of a particular skill) by the student, performance assessment is both time-consuming and expensive. Care needs to be taken to ensure that assessment tasks are designed to reflect the desired skills and knowledge. Performance assessment also requires the professional judgment of a marker (rater/examiner) to provide a measurable result. The requirement for markers adds to the cost and also reduces the reliability of the assessment, since there will inevitably be differences of opinion between markers about students' work. Wood (1991, Ch 5), for example, reviews a number of studies that show that the correlation between essay marks awarded completely independently by one marker and another rarely rises above 0.6. However, the principal advantage of performance assessment is that it relates to curricular goals and supports classroom teaching, and so tries to make positive use of teaching to the test (Wiggins, 1989). A sub-category of performance assessment, labelled authentic assessment, involves teachers conducting assessment of actual classroom activities as they are carried out during the course of normal teaching. Such authentic assessment is normally seen as a part of formative assessment, but it can also form part of a summative assessment model.

2.4 Assessment and learning

Standardized objective achievement tests were developed on the basis of behaviourist theory, which claims that learning derives from the establishment of a large number of stimulus-response relationships (Resnick and Resnick, 1992). Complex knowledge and skills can be broken down into discrete, de-contextualized building blocks. Learning is seen as linear and sequential.

Performance assessment, if properly designed and conducted, takes due account of a more modern understanding of the nature of learning. Research in the areas of cognitive psychology and learning (see, for example, Resnick, 1989; Shepard, 1991) reveals that students best achieve deep understanding through interpretation and construction of knowledge and skills, by extending their existing knowledge structures (or schema). Learning often takes place in an irregular fashion, not following a logical sequence from simple to complex. This constructivist approach acknowledges a more 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 modern constructivist learning theory (Black, 1999; Shepard, 1992; Wood, 1998; 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 is a misleading distinction that 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 the teaching and learning of a subject. This level of integration goes beyond what is implied by the term "backwash", and can be expressed in the paradigm below (from Furst's paradigm, 1958, in Frith and Macintosh, 1984).



[1] Different authors have used differing interpretations of the term "performance assessment" to suit their own purposes. For this booklet, the term is used in its widest sense to cover all forms of assessment in which students are assessed directly on how they demonstrate achievement of the learning objectives of a course of study.

2.5 Norm-referencing and criterion-referencing

The terms norm-referencing and criterion-referencing represent another dimension along which assessment systems can differ, a dimension that reflects the means by which outcomes of an assessment process are reported on a consistent scale over a period of time or over different events. One method, generally associated with standardized tests, is to trial the test on a typical sample of students, 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 student taking the same test. This is called norm-referencing, and the process of deriving a standard distribution of scores from the initial trial is called norming. 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 students can vary from this normal distribution. In principle, norm-referenced tests can be used to report changes in student performance over time; they do not only report a student's score in comparison to his/her immediate peers. Over the years, with repeated use of a given test, results can vary upwards or downwards. Cannell (1988), for example, found that elementary school pupils in all 50 states of the USA were recording above-average scores on nationally normed basic-skills tests. This finding generated a number of possible explanations for the effect, of which improved teaching and learning was one of the less well supported.

Despite the potential for norm-referenced tests to measure changes in standard of achievement over time, the term norm-referencing is now also commonly applied to those assessment systems that impose the same fixed distribution of results on repeated occasions. In such systems, each student's score, or grade, is in effect only a measure of their position in the rank order of all those students who took the assessment at that time. This says less about what they have achieved than it does about who has done better or worse than they have, which may of course be quite appropriate for certain quota-based selection instruments.

The notion of criterion-referenced assessment was first put forward by Glaser (1963). Criterion-referenced assessment represented a significant shift away from aptitude testing, towards an emphasis on measuring student achievement "with respect to a well defined behavioural domain" (Popham, 1978). It also led to a further influential approach in the USA, called measurement-driven instruction (Popham, 1987), which harnesses the backwash effect of high-stakes criterion-referenced testing to influence the instructional programme leading up to the test. However, it should be noted that, in its proper form as originally defined, criterion-referenced assessment is still based on behaviourist learning theory and involves standardized objective tests. The tests are built around student performance in a restricted domain of learning (such as adding simple fractions) and assume a discrete and hierarchical set of skills. The principal distinguishing features of criterion-referenced tests are that:

1. criterion-referenced test items are selected to represent discrete units of student learning, and
2. the outcome of the test depends on whether the student has reached a theoretically pre-determined cut-off score rather than how the student's score compares to a pre-determined distribution of performance.

The outcome of a traditional criterion-referenced test is that mastery of the relevant domain has either been shown or not shown. Criterion-referenced tests and norm-referenced tests differ more in the analysis and interpretation of student responses than they do in the kind of questions set.

Outside the USA the terms norm-referenced and criterion-referenced are often used loosely to represent very different philosophies of constructing assessment instruments, reporting results and measuring standards of achievement. However, the distinction between the two approaches should not be overestimated. The initial norming of norm-referenced tests can set the standard by which future students are measured, and the standard of achievement required to demonstrate mastery on a criterion-referenced test is often decided upon in relation to expected performance across the population as a whole—norms can represent standards, and criteria are established on the basis of normative data.

Attempts have been made to introduce formal criterion-referencing into the UK assessment system in the context of the GCSE examinations, which were introduced in 1986 (Orr and Nuttall, 1983; SEC, 1984; Kingdon and Stobart, 1987). Criterion-referencing was also attempted in UK national curriculum testing, with its associated attainment targets (Brown, 1988). These attempts were not successful. Criterion-referenced assessment, in its strictest sense, is only suitable for relatively straightforward, easily defined, uniform kinds of task, and is not suited to measuring performance in more complex subject areas. To do so requires either a burdensome and complex system of assessment or loosely expressed and general criteria, allowing for variation in student performance across different parts of the overall assessment, which does not suit the original model. A strict criterion-referenced model would demand mastery in every aspect of a subject before a student could be said to have "passed", and would report student achievement according to their worst aspect. In the UK and some other countries, a more generalized approach has been adopted, under a variety of labels such as standards-referenced assessment (Sadler, 1987), standards-based assessment, outcomes-based assessment, construct-referenced assessment (Black, 1998) or criterion-related assessment.

2.6 Validity and reliability—an overview

According to the standard definition, the validity of an assessment is the extent to which it actually measures what it is stated to measure. The term reliability is used to define the accuracy of measurement resulting from an assessment, and how likely it is that the same result would be produced in slightly different circumstances. An assessment is reliable if a student would gain the same result were he/she to repeat the assessment on different occasions, and also give the same result if the assessment were marked by different markers. Validity and reliability are widely regarded as essential characteristics of any assessment system, particularly a high-stakes one where the outcome is of great importance to the student or the teacher. Both characteristics are in fact multi-modal. There are different types of validity and also several different kinds of reliability.

The terms validity and reliability, and the approaches to their measurement, were derived very much from a psychometric background. The construct, or aptitude, being measured is defined by the test itself and uniformity is ensured by weeding out questions that produce erratic responses by test takers. The construction of standardized tests as large collections of items, which all behave in a similar and predictable fashion, almost inevitably leads to high measures of reliability and confident statements about validity, because the whole test can be shown to relate to one aptitude or skill area, which can readily be given an appropriate label. In psychometric testing, reliability and validity (in a restricted sense) become inextricably intertwined. Validity is seen largely as identifying a single construct measured by a given test, and reliability is seen largely as how consistently the different items in the test behave, in terms of the correlation of student responses given to these different items.

In the context of performance assessment, the concepts of validity and reliability take on slightly different interpretations. Validity is broadened to include the social and educational consequences of conducting assessments, acknowledging the interaction between assessment and teaching. Performance assessment instruments are rarely designed to measure only one narrowly defined uniform construct, and it is accepted that reading and writing skills, for example, should not be completely divorced in assessment from subject skills. The interpretation of assessment results and the uses to which they are put become part of the validity of the assessment. The emphasis on precision and accuracy of measurement is lessened. It is accepted that high levels of technical reliability are not achievable in an examination system (for example, Wood, 1991; Satterley, 1994).

However, given the broad and complex educational skills and achievements a performance assessment is usually attempting to address, particularly in relation to students at the end of their secondary education, there is considerable doubt as to the worth of a concept such as a "true score" for each student, that assessment processes should be trying to identify. The different skill areas and wide range of possible contextual frameworks in which skill and

knowledge could be demonstrated are so diverse that precise, comprehensive measurement is not possible.

In addition, student capability across this potential range of achievements is not a static, invariant quantity, such as each student's height would be, but is something more dynamic and variable in nature. Precision of measurement by any particular assessment instrument, even if it could be achieved, would not be meaningful in that it would represent a student's achievement only on that particular collection of tasks and only at that particular time. Student variability is a major factor in making high reliability impossible. A degree of approximation must be accepted, although this should not be taken to mean that assessment organizations can afford to ignore a strong requirement to make results as dependable as possible, at the level at which they are reported.

Gipps (1994) has argued strongly for a paradigm shift in order to re-conceptualize reliability in the context of educational assessment, as opposed to psychometric testing. She states, (p167):

> "Assessment is not an exact science, and we must stop presenting it as such. This is of course part of the post-modern condition—a suspension of belief in the absolute status of 'scientific' knowledge. The modernist stance suggests that it is possible to be a disinterested observer, while the post-modernist stance indicates that such detachment is not possible…The constructivist paradigm does not accept that reality is fixed and independent of the observer; rather reality is constructed by the observer, thus there are multiple constructions of reality. This paradigm would then deny the existence of such a thing as a 'true score'."

To summarize, validity and reliability were originally given restricted definitions, enabling them to be bound together and fit comfortably within the theoretical framework of psychometric testing. Ability and attainment are each seen as a unified property of an individual, which can be accurately and precisely measured. In the broader context of performance assessment, or educational assessment, validity and reliability take on wider meanings and a conflict between them has been recognized (for example, Harlen, 1994). Improvements in levels of construct validity may often only be achieved at the expense of reliability, and vice versa. Good assessment comes from achieving a satisfactory compromise, and the nature of the balance between reliability and validity will depend on the context and purpose of a particular assessment system. Gipps (1994) refers to the concept of "dependability", defined as the optimum mix of validity and reliability in relation to a given type of assessment. In formative assessment, pre-eminence can be given to validity, while in summative assessment equal attention must be paid to both validity and reliability.

2.7 Bias

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 test items/assessment tasks themselves, or from the marking of a performance assessment. In the latter case bias becomes an issue of marking reliability.

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 student population ("differential item functioning" or DIF) may be regarded as biased and removed from the test. The student 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 student 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 have great potential. 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.

Decisions about whether certain test items are biased or legitimate should be based on how each item can be explicitly linked to the underlying construct and what the possible factors for introducing bias might be, rather than on purely statistical grounds based on item calibration. Goldstein (1996a) 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 (1998, p50) 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, mechanical toys and certain sports favour boys, dolls and domestic work favour girls).
- Essay questions on impersonal topics favour boys, while those involving human relations favour girls.
- Multiple-choice questions favour boys.
- Coursework/project work components of assessment favour girls.
- Some questions may be intelligible only within certain cultures, for example a question about elderly people living on their own might be quite alien to some cultures, or a question involving a typical male or female role from one culture may appear very out of place in another.
- A question using language or conventions of one social class would favour students from that class.

There is a substantial body of research into such differences, which are well established (see for example, Wood, 1991, Ch 14). The manner in which assessment designers should respond to such differences is not always quite so clear. Assessment instruments 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 students who have different kinds of learning style, 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 student 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 students. Any or all of these could significantly affect an individual student'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 student 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.

A further aspect of bias that must be countered is the potential for an assessment task to discriminate unfairly against students with special educational needs such as dyslexia, attention deficit disorder or impaired vision. The conditions under which assessment tasks are taken should make appropriate allowances for such students, so that they can demonstrate their level of educational achievement on equal terms with other students.

Bias arising in marking can occur for a number of reasons, such as personal attitude to neatness of student handwriting (for example, Hughes *et al*, 1983), preferential treatment for student 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.

3. Diploma Programme assessment—aims and approaches

This section describes the philosophical approach adopted in the context of the formal assessment used for the IB Diploma Programme (DP). This includes the manner in which account is taken of basic principles of assessment (described in section 2), and provides the foundation on which the structures and processes (see sections 4 and 5) are constructed. By its very nature, formal DP assessment is summative assessment, designed to record student achievement at, or towards the end of, the course of study. It should be noted, however, that many of the assessment instruments can also be used formatively during the course of teaching and learning, and teachers are encouraged to do this. This is particularly true of the internal assessment tasks (see sections 4.2 and 5.3).

Assessment of the DP is high-stakes, criterion-related performance assessment. It is based on the following aims, which are elaborated in the remainder of this section.

1. DP assessment should support the curricular and philosophical goals of the programme, through the encouragement of good classroom practice and appropriate student learning.
2. The published results of DP assessment (that is, subject grades) must have a sufficiently high level of reliability, appropriate to a high-stakes university entrance qualification.
3. DP assessment must reflect the international-mindedness of the programme wherever possible, must avoid cultural bias, and must make appropriate allowance for students working in their second language.
4. DP assessment must pay appropriate attention to the higher-order cognitive skills (synthesis, reflection, evaluation, critical thinking) as well as the more fundamental cognitive skills (knowledge, understanding and application).
5. Assessment for each subject must include a suitable range of tasks and instruments/ components that ensure all objectives for the subject are assessed.
6. The principal means of assessing student achievement and determining subject grades should be the professional judgment of experienced senior examiners, supported by statistical information.

3.1 Support for curricular goals

Although the list above is not presented in order of descending priorities, and some of the aims are interrelated, it is clear that the single most important aim of DP assessment is that it should support and encourage appropriate student learning. This is the feature most valued by the users of the diploma qualification, mainly institutions of higher education (IBO, 2003a), and by the schools and students themselves. Absolute reliability of assessment results, though highly important in its own right, cannot take priority over student learning.

There is a very real conflict in assessment design between techniques that can give the most accurate and reliable measures of certain aspects of student achievement, and techniques that measure and encourage the most desirable educational achievements of students. This dilemma was very well acknowledged by Alec Peterson (1971), who described the early development of DP assessment as follows:

> "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 [is] sufficiently reliable to assure pupils, parents and 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."

The strong impact of high-stakes assessment on teaching and learning must 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). If the aim of the DP is to achieve the development of students who are "inquiring, knowledgeable and caring" and who become "active, compassionate and lifelong learners" (IBO mission statement), then these characteristics should be reflected in the assessment system. It is an inevitable fact that what is not assessed is not so highly valued and may even be overlooked altogether. The aspirations expressed in the mission statement must be supported by the assessment system.

The desired personal characteristics of students, expressed in the IBO 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 students' 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 largely achieved through the creativity, action, service (CAS) requirement, though there are a number of references to ethical working practices elsewhere in the assessment system.

In terms of assessment principles, DP assessment places a strong emphasis on consequential validity (see appendix A.1), with an awareness that the manner in which assessment is conducted will have a major impact on how the DP is taught within schools. This impact has been deliberately enhanced in recent years by the provision of increasing amounts of feedback to schools and teachers, about the performance of their students on the assessments and ways in which that performance could be improved.

Comprehensive research studies on the predictive validity of DP results have yet to be conducted, but the small number of informal studies carried out and the substantial amount of anecdotal evidence suggest that the predictive validity of diploma results is high. The assessment model (collection of assessment instruments) applied to each subject is designed to be broadly based, including a variety of types of evidence, both to ensure construct validity and to improve the generalizability of the results as much as possible.

3.2 Reliability of results

Although reliability of results at the level of subject grades must be a priority for a high-stakes assessment system, absolute precision of measurement to within a mark on every task undertaken by a student is not possible or even necessary. The aim is to have at least 95% confidence that any final subject grade is "correct". A correct result in this sense is a result that would be confirmed by subsequent re-marking of candidate work by the most senior examiners. This is a reasonable target for a system heavily dependent on qualitative judgment rather than technical measurement, and quality control systems are utilized to ensure this target is met. The assessment model used to generate a subject result consists of a variety of tasks, taken in different contexts on different occasions. This helps to reduce threats to reliability posed by a single assessment task given in one particular context. Internal consistency measures of reliability (see appendix A.2) are not considered appropriate because each component (assessment instrument) may deliberately contain varied forms of task, or sometimes a small number of tasks. Parallel-forms reliability (see appendix A.2), relating to the new examination papers created for each session of assessment, is not essential at the level of student marks awarded. It is accepted that marks may be slightly harder or easier to achieve on different instances of the "same" assessment instrument. The burden of reliability falls on determining grades that consistently represent the same standard of achievement (see section 5.6). A strong emphasis is placed on ensuring marker reliability through the use of detailed markschemes, assessment criteria and moderation procedures (see sections 5.4 and 5.5) and reducing marker bias. Marking reliability, as it affects the final subject grade, is also improved by the use of different markers to mark different parts of a student's total assessed work for a subject.

Great care is also taken to ensure grading reliability, through the application of consistent standards supported by statistical background data, in determining grade boundaries. Grade standards are documented and exemplified, and judgments made about grade boundaries are checked by a number of statistical indicators.

In summary, as the recording of student achievement progresses through different levels from component mark through component grade to subject grade, the reliability of reporting increases. High reliability is achieved at the level of final subject grade.

3.3 International and intercultural dimensions

The DP is studied by students in over 100 countries and with many more nationalities. As well as the academic aims of the programme, the IBO intends that, through following the programme, 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" (IBO mission statement). There is, therefore, both an international context and an intercultural purpose to teaching, both of which must be reflected in the assessment. A major factor in this is language. Diploma Programme assessments are conducted in English, French and Spanish. Examination papers are normally prepared in English and translated into French and Spanish. Comparability of demand across the three languages is aided by sensitive translation, which can give rise, when necessary, to amendments to the original English version of the examination paper. If, on subsequent consideration of actual performance, it seems that some slight advantage or disadvantage has been unintentionally given to students working in one particular language, then a corresponding adjustment to marks is made. Many senior examiners are bilingual or trilingual.

The DP offers a wide range of second-language courses, for different levels of proficiency, and additionally guarantees that any student can follow a literature course in their own best language, provided there is a sufficient body of literary works in that language to form an adequate basis for study.

Beyond the issue of language, a significant cross-cultural dimension is included in many DP subjects and their assessment. The following are some examples.

- In the literature course (language A1), students must study some works that were originally written in a language different from the one being followed for their course. Students must write assignments, including a cross-cultural perspective, on these works.
- In second-language courses (for example, language B), the language should be studied in a strong cultural and practical setting. Assessment of language use includes awareness of the cultural setting.
- The history course includes a compulsory section on world history and develops an international perspective on historical explanation.
- The economics course includes substantial subject content on international economics and development economics; students are expected to understand economic theory and application from different national and cultural perspectives.
- In the music course, students must carry out an investigation into the relationship between two musical genres from different cultures.

In some other subjects, the issue of cultural variety is dealt with more through a tolerance of different cultural emphasis made possible by means of the subject syllabus structure. 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 to a certain extent to suit particular cultural traditions of teaching the subject.

The DP courses of study are thus tolerant of cultural variants as well as encouraging of cultural tolerance. This poses assessment problems in terms of maintaining comparability across the optional approaches that are permitted for part of many subjects. Assessment

comparability within a subject is always compromised when there are choices of question, options or very open-ended assessment tasks. However, the increased utility and assessment validity provided by these structures to students in varied cultural settings in different parts of the world override such concerns. The crucial factor is that students follow courses of study that are appropriate to their own cultural context. The same biology course, for example, would not suit all. Attempts should of course be made to ensure as high a degree of comparability as possible. Information on comparability can be gained through analysis of student performance on core aspects of the assessment in relation to performance on optional aspects of the assessment.

There is more to interculturalism 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. There is no intention to include any kind of psychological profiling in formal DP assessment. Instead, student contributions by their actions, which reflect their values and attitudes, are part of the creativity, action, service (CAS) requirement (see section 4.1). There is no scale of achievement or grading associated with CAS, but schools must authenticate the satisfactory participation of each DP student. Without this authentication, students cannot be awarded the diploma. CAS therefore has a significant impact on the overall outcome of DP assessment. Practical interculturalism is encouraged in the assessed work for arts subjects, and possibilities for practical involvement may also arise in the internal assessment work for some other subjects (see section 4.2).

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 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 DP 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 students, 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 utilize more open-ended assessment questions and tasks that allow students 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 DP assessment philosophy (see section 3.4).

Even with the application of both these methods, students 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 Diploma Programme and assessment philosophy, in that one of the aims of the programme 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.

Earlier in this booklet (see section 2.7) research was quoted indicating the equity issues surrounding different assessment formats. Much of the research in this area has related to gender (Wood, 1991, Ch 14; Gipps and Murphy, 1994). In an international setting, different assessment formats may also give rise to inequity on a cultural basis, related to past educational experience. Some DP students may have previously been exposed only to multiple-choice tests, others only to lengthy essays, or oral interviews, or portfolios as modes

of assessment. This is yet another reason for including as wide a range of formats as possible within the DP assessment system (see section 3.5), so that all students and their teachers meet some formats with which they are familiar and more comfortable, and other formats with which they are less familiar.

Two further issues of cultural equity in DP assessment relate to examinations taken in a second language and the role of group work.

A significant proportion of DP 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 languages in which DP 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. Sentences should be short, with simple wording and sentence structure used wherever possible. However, subject-specific terminology should not be avoided. Additionally, tolerance must be shown towards errors in spelling and grammar when marking is carried out, except in languages examinations. As long as the meaning and communication are clear, no penalty should be applied and full marks should be available.

Finally, DP 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. Students 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.

Group work poses significant problems for assessment, in reliably identifying who has contributed what and who has benefited (or suffered) unfairly from the work of others. However, in the interests of assessment validity as well as cultural equity, more must be done to include cooperative group working in many assessment systems. Diploma Programme assessment does include a limited element of cooperative group work. In all the science courses, students must participate in an interdisciplinary project, which by its nature requires group work. One of the assessment criteria applied to practical work in the sciences relates to how well a student engages in team work, and the interdisciplinary project is a suitable context in which teachers can assess this.

In a slightly different vein, the music course has a performance component, which may optionally be a group performance. In this case each student in the group is awarded the same mark, according to the whole group's performance. It is up to each group member to try to ensure the best possible performance by the group as an ensemble. In the theory of knowledge course, students have to make a presentation to the rest of the class on a knowledge issue. The presentation must form an integral part of the teaching of the course. Presentations can be individual or by a group, but if by a group the teacher awards marks on an individual basis to students according to their contribution.

These instances form only a small part of the overall DP assessment, and it would be true to say that group work is still considerably under-represented in the overall structure.

3.4 Higher-order cognitive skills

It is often stated that we now live in a knowledge society. This should not be taken to mean that the acquisition and retention of factual information is of prime importance. The explosion of information in recent times makes it impossible for individuals to achieve mastery of knowledge in many areas. The more valued academic skills of today are in accessing, ordering, sifting, synthesizing and evaluating information, and creatively constructing knowledge. As the rate of development and change in many societies increases, learning to

learn becomes a more valuable skill than just learning knowledge and concepts. A similar point was made by Peterson (2003), the one person above all others who shaped the educational philosophy of the IBO. He stated that "what matters is not the absorption and regurgitation either of facts or of predigested 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".

Such considerations must have a major influence on assessment if it is to retain validity. Assessment relating only to the recall of knowledge, concepts and routine techniques is no longer sufficient. If the skills expected of today's students are changing, or rather expanding to include a greater diversity, then assessment instruments should do likewise. The DP assessment system deliberately attempts to give significant attention to the so-called "higher-order" cognitive skills (Bloom *et al*, 1956). 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 assessing student achievement in these areas, and because the outcomes of such activity cannot be tightly prescribed, such assessments must be comparatively unstructured and open ended. This inevitably raises further concerns about marking reliability of the outcomes, when there may be many diverse but correct responses.

Nevertheless, the demands of construct validity, when the construct includes sophisticated and complex productive skills, cannot be ignored. One of the long-standing aims of the DP has been to develop students who are critical thinkers (Hill, 2002). The objectives of the subjects in the programme invariably include significant representation of Bloom's higher-order skills, and so these must form a significant part of the construct to be assessed. In practice, it is not always simple to separate out the skill levels, to determine which skill level a student might use to respond to a given question, or even to say which skill level represents the greater educational demand in a given context, but this does not excuse assessment systems from the need to address the full range. To this end, DP assessment must include substantial tasks that require students to reflect on their knowledge and construct extended pieces of work in response to the task set.

3.5 Range of assessment tasks and assessment instruments (components)

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 assessment tasks. An assessment instrument/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 student may carry out only one task out of a number of choices available for a component. A given component is marked by a single marker for each student's work in the DP assessment system.

There are a number of reasons why a wide variety of types of assessment task and component is used in relation to the DP. 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". Second, a variety of assessment techniques helps to reduce the potential for inequity in assessment, as discussed in section 2.8 (see also, Linn, 1992; Brown, 2002). There are also theoretical considerations, relating to fitness for purpose, that require a varied approach to assessment. The range of components and the set of tasks within them ensure that, taken across the assessment model for a whole subject (see section 4.2), student achievement against all the objectives for that subject is adequately represented.

The construct being assessed for each subject is defined by the objectives given at the beginning of each subject guide. The nature of what is to be assessed is thus precisely defined for students, teachers, parents/carers and examiners. Because the objectives can represent a wide variety of skill types, the assessment tasks and components may likewise vary considerably within and across subjects. Examples for two different subjects will illustrate the point. First, consider three of the nine objectives for the literature course, language A1 at higher level (see section 4.1). These are as follows (IBO, 1999, p6):

> "…candidates will be expected to demonstrate:
>
> - an ability to engage in independent literary criticism in a manner which reveals a personal response to literature
> - an ability to express ideas with clarity, coherence, conciseness, precision and fluency in both written and oral communication
> - an appreciation of the similarities and differences between literary works from different ages and/or cultures."

Expressed in this way, these and the other six objectives give a very clear idea both of the kind of skills that need to be taught and of the kind of assessment tasks that must be employed to allow students to demonstrate these skills. Student achievement against the first objective is demonstrated through an oral commentary on an unanticipated extract from a studied work, and through a two-hour essay written under examination conditions on two of the works studied. The second objective is addressed across all the assessment components (two examination papers, two world literature assignments and an oral commentary and presentation). The third objective is addressed through a comparative study, carried out over a period of time rather than under examination conditions, of at least two works from the world literature section of the course.

The generic nature of the objectives, and the higher-order skills expressed in them, provide clear indications of the possible assessment format. Although the tasks are open, and clearly must be marked more by professional judgment than analytical point scoring, teachers and students are given substantial guidance on the parameters of the task, and examiners (markers) are given detailed sets of assessment criteria by which to mark the work.

All the science courses have the same set of objectives, which are as follows (IBO, 2001b, p7):

> "1. Demonstrate an understanding of:
> a) scientific facts and concepts
> b) scientific methods and techniques
> c) scientific terminology
> d) methods of presenting scientific information.
>
> 2. Apply and use:
> a) scientific facts and concepts
> b) scientific methods and techniques
> c) scientific terminology to communicate effectively
> d) appropriate methods to present scientific information.
>
> 3. Construct, analyse and evaluate:
> a) hypotheses, research questions and predictions
> b) scientific methods and techniques
> c) scientific explanations.
>
> 4. Demonstrate the personal skills of cooperation, perseverance and responsibility appropriate for effective scientific investigation and problem solving.
>
> 5. Demonstrate the manipulative skills necessary to carry out scientific investigations with precision and safety."

The structure of these objectives follows Bloom's taxonomy (Bloom *et al*, 1956) fairly closely. All science subjects are assessed through four components: three examination papers and practical laboratory work, which is marked by the classroom teacher. The first examination paper is a multiple-choice test, designed to give broad coverage of course content, assessing objectives 1 and 2. The second examination paper consists of a data-analysis question, some short-answer questions and one extended-response question (two at higher level) based on the core course content followed by all students. Paper 3 consists of short-answer questions on the particular options students have chosen to study. In both papers 2 and 3, questions are designed to give equal weighting to objectives 1 and 2 on the one hand, and objective 3 on the other. The practical work includes assessment of all five objectives. For science subjects, the structure of the examination papers ensures balanced and appropriate coverage of course content.

The above examples illustrate how construct validity (see appendix A.1) is established and maintained throughout the DP. The course objectives for each subject define the construct, provide a framework for course content, significantly influence subject teaching, determine what assessment tasks and instruments should be used, and also provide strong guidance to markers on what characteristics of student work should be given credit. Objectives are usually defined in terms of skills. The degree to which objectives are linked to prescribed course content will vary from subject to subject (low direct linkage in language A1, for example, and high direct linkage in science subjects and mathematics), and this also influences the format of assessment components and the tasks that make them up.

3.6 The role of professional judgment

The complex, higher-level educational skills that form the focus of DP assessment do not lend themselves readily to atomized, deconstructed marking. Student responses to many assessment tasks can be highly varied, with many equally valid and correct forms of response. Often, it is not possible to give precise guidance to markers on exactly what each mark should be awarded for, nor is this desirable. Section 2.4 referred to developments in learning theory, which suggest that complex knowledge and skills should not be taught by breaking them down into small, decontextualized building blocks. The same principle applies to the tasks designed to assess learning and to the way in which markers should assess student responses. The very complexity of the skills being assessed denies the possibility of mechanistic marking. There are some subject areas, for example, mathematics and the sciences, where analytical marking approaches are the general rule, but even here it is not possible to prescribe exactly what response students should produce, and markers must be constantly aware of the need to give credit for alternative valid responses.

Much then depends on the professional judgment of markers, and particularly on the professional expertise of the senior examiners who monitor the work of all markers (markers here include teachers as well as examiners). This represents a strong challenge to the reliability and integrity of the assessment system, but a challenge that must be met. Considerable effort has to be expended on providing instruction, guidance and support material to all markers, and on developing quality control systems to check marking standards (see sections 5.4 and 5.5). IBO staff work closely with senior examiners in the development of these support and quality control mechanisms.

The reliance on professional judgment is present in many standards-referenced, or criterion-related, systems and in many competency-based systems, which are often used in vocational qualifications. Gardner (1999) makes a strong contrast between assessment through standardized objective testing at one extreme and assessment through apprenticeship methods, where learning takes place in a practical work context, at the other. An apprentice's progress is judged by the master, who has a close knowledge of the work processes of the apprentice as well as a broad view of work output in a variety of contexts. Gardner strongly counters accusations that such a system is too subjective, arguing that sufficient reliability can be established and that in any case, so-called objective testing is, in practice, highly biased towards students with a certain blend of linguistic and logical intelligences.

The formal assessment system adopted by the IBO for the DP includes elements from both extremes. There are some highly formalized assessment tasks, including multiple-choice tests, but there are also many more substantial open-ended tasks, and a focus on process through teacher marking of student projects and practical work. This range of student performances is reduced to a final subject grade (on a scale of 1 to 7) according to grade descriptors, which represent the standards for each subject. Such standards may exist on paper in generic form, and be reinforced by written exemplar material, but in the final analysis the complexity and variety of information that must be synthesized in order to arrive at a judgment requires an interpretation of standards resident in the minds of the experienced senior examining team. The senior examiners and IBO staff do refer to statistical data to verify their judgments, but the primary decisions in marking and grading are based on the judgment of student performance against the expected standards.

4. Diploma Programme assessment structures

4.1 The Diploma Programme curriculum

The Diploma Programme (DP) is a two-year course of study for students aged 16 to 19. It offers a broad and balanced curriculum, which is a deliberate compromise between the early specialization preferred in some national systems and the breadth favoured in others. The curriculum model can be represented as a hexagon, with six academic areas surrounding the core. Subjects are studied concurrently and students are required to follow a range of subjects representing all the major disciplines.

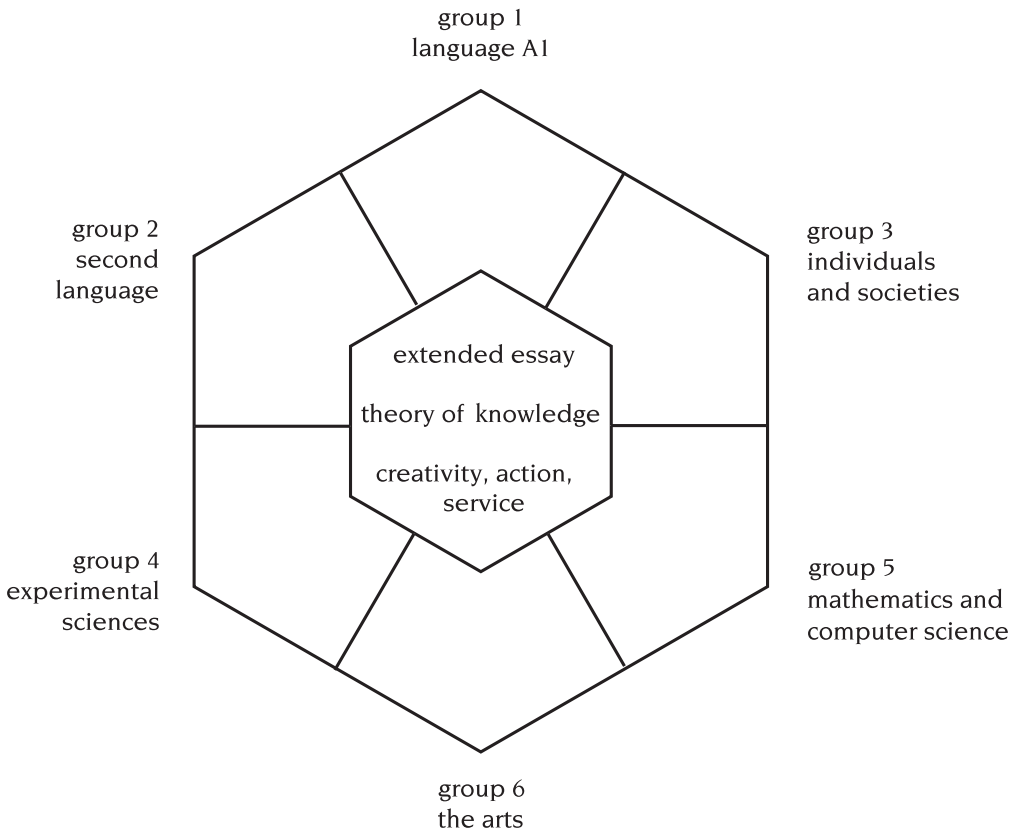


Figure 1: *The Diploma Programme hexagon curriculum structure.*

Students studying for the full diploma are required to select one subject from each of groups 1 to 5. A sixth subject is chosen either from group 6, or as a second subject from one of the other groups. Three subjects (occasionally four) are taken at higher level (HL) and three subjects (occasionally two) are taken at standard level (SL). The recommended teaching time is 240 hours for an HL course and 150 hours for an SL course. This allows students to go into more depth in their preferred subject areas, while requiring them to continue to study in other areas. SL courses are often subsets of HL courses in the same subject.

If students are unable to study the programme in full, they can follow fewer courses, for which they will receive individual certificate results.

Group 1: language A1
Group 1 consists of literature courses in a student's first language. The courses introduce students to literature from a variety of periods, genres and styles. Students refine their skills in writing, speaking and analysis, and learn techniques of literary criticism. The courses help students maintain strong ties to their own culture while giving them an international perspective through the study of literature from around the world.

Group 2: second language (language *ab initio*, language B, language A2, classical languages)
The acquisition of a second language carries great importance in the DP. Students learn to understand and use the language, and gain insights into the cultures of the countries where the language is spoken. This subject group includes courses for beginners (language *ab initio*, classical languages), second-language learners with previous experience with the language (language B), and bilingual students with a high level of fluency (language A2).

Group 3: individuals and societies
This group includes nine subjects: economics, geography, history, philosophy, psychology, social and cultural anthropology, business and management, Islamic history, and information technology in a global society. By studying human experience and behaviour, as well as economic and social environments and institutions, students gain an appreciation of diverse perspectives and values. They learn to analyse concepts and theories, and to use quantitative and qualitative methods of data collection and analysis.

Group 4: experimental sciences
The sciences offered in this group are biology, chemistry, physics, environmental systems and design technology. Students become familiar with the body of knowledge, methods and techniques that characterize science and technology, and learn practical laboratory skills.

Group 5: mathematics and computer science
This group includes courses designed for a range of abilities and interests. Some are aimed at students who wish to study mathematics in depth, while others are for those who need mathematics to enhance their understanding of other subjects. The courses seek to provide students with mathematical knowledge and principles. They help students develop logical and creative thinking in mathematics and use abstraction and generalization to reach conclusions. All students must follow a course in mathematics, and may also elect to study computer science.

Group 6: the arts
The arts group includes visual arts, music and theatre arts. The emphasis is on making art. That is, students gain an understanding of the arts and learn to express themselves artistically by creating, producing or performing works of art. In addition, they explore art forms from different cultures throughout the world.

Core requirements
At the heart of the DP are three requirements that students must fulfill in addition to their work in six subjects.

- **Theory of knowledge**
  One of the most important elements of the DP is the theory of knowledge course, which challenges students to question the bases of knowledge—to reflect critically on how they know what they believe to be facts or the truth. It consists almost entirely of exploring questions about different sources of knowledge (perception, language, emotion, reason) and different kinds of knowledge (scientific, artistic, mathematical, historical), such as:
  - Do we construct reality or do we recognize it?
  - Does knowledge always require some kind of rational basis? Is there any kind of knowledge that can be attained solely through emotion?
  - Is scientific knowledge progressive; has it always grown? Can we reach a point where everything important in a scientific sense is known?

- **Creativity, action, service (CAS)**
  Another important element of the DP is creativity, action, service (CAS). To fulfill this requirement, students must take part in artistic activities (creative); sports, expeditions or local or international projects (action); and community or social service projects (service). Participation in CAS raises students' awareness of community needs and gives them an opportunity to apply what they have learned in the classroom to address these needs. It also gives them confidence in their ability to bring about change. The projects must have tangible results and offer real benefits to others. Reflection on their experience is also an important part of student involvement in CAS.
- **The extended essay**
  An extended essay, of at most 4,000 words, offers students an opportunity to conduct an in-depth study of a topic of special interest. The experience and skills gained in carrying out independent research and producing a structured, substantial piece of writing provide excellent preparation for independent study at university level.

The DP's curricular structure defines the framework in which assessment must operate. Individual assessment models are constructed for each subject at both HL and SL, for theory of knowledge (TOK) and for the extended essay. Two examination sessions are held each year, in May and November, with results being released in early July and early January respectively. The published results are made up of subject grades, which equate to diploma points, in the range from 1 (lowest) to 7 (highest) at HL and at SL, and grades from E (lowest) to A (highest) for TOK and the extended essay. A matrix table converts the combined letter grades for TOK and the extended essay into a points score from 0 to 3. CAS does not contribute to the points total, but authenticated participation in CAS is a requirement without which the diploma cannot be awarded.

Thus, the maximum possible points total for a DP student is 45 (6×7, plus 3). A student gaining 24 points or more, subject to certain conditions relating to the distribution of points across subjects, will be awarded the diploma. The policy of making the same number of points available for both HL and SL courses, despite the difference in workload and achievement at the two levels, is a deliberate one, encouraging students to regard their SL courses as equally important to their HL courses. Students are encouraged to achieve their best across all disciplines and are appropriately rewarded for doing so.

4.2 Assessment models and the role of internal assessment

In order to provide for the formal assessment of the programme, assessment models are devised for each level (HL and SL) of each subject and also for each of the core requirements except CAS. Each model will consist of a few different assessment components, usually including a range of different tasks. Assessment models are revised as part of the curriculum review process for each subject. Subjects are normally reviewed every seven years by a review group consisting of teachers, examiners, IBO staff and external consultants. The curriculum review process is a consultative one with proposals being circulated to authorized DP schools for comment as they are developed. Recommendations and proposals from the curriculum review groups are also submitted to the diploma review committee for consideration. This committee has responsibility for the overall academic quality of the courses that make up the DP and approves proposed syllabuses and assessment models. The committee is particularly concerned with:

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

In the context of assessment, both the curriculum review groups and the diploma review committee (DRC) refer to the DP assessment policy (see appendix B), which defines the parameters under which assessment models can be developed.

The majority of assessment components are examination papers, made up of a wide variety of question types to suit the requirements of the subject. Question types include multiple-choice questions (used in only a small number of papers), short-response questions, structured problem-solving questions, open-ended problem-solving questions, essay questions, data-analysis questions, case studies and commentaries on supplied texts. The examinations are taken under controlled conditions, with student responses being marked externally by independent examiners.

There are other tasks/components undertaken by students, with the guidance of their teachers, over an extended period, which are also externally marked by examiners. These include language A1 world literature assignments, language A2 written tasks, music investigations, theory of knowledge essays and extended essays. The assessment focus for all of these is on the quality of a finished written product, which makes them suitable for external assessment.

A third type of assessment component is internal assessment, that is, student work marked by the teacher with this marking subject to external moderation (see section 5.5). The DP assessment policy allows for the existence of an internally assessed component for any course where it is considered appropriate. Very few courses do not have an internally assessed component. Internal assessment allows for components/tasks to be included in the assessment model that provide evidence of student achievement against objectives that do not lend themselves to external examination. This particularly relates to process skills, as may be demonstrated in such activities as project work, fieldwork, laboratory practical work and mathematical investigations. Although workbooks and portfolios may be used to record process skills, these in themselves do not make suitable tools for external assessment. However, they do provide a means by which moderators (external examiners) can check that the standard of teachers' marking is appropriate. Internal assessment is also used for oral work in languages courses, which allows teachers to choose the most appropriate opportunity to carry out formally assessed oral work and also to provide a supportive environment for it.

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 immediately outside. International schools, whose students often have a different cultural background from the one 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 electronic 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 some significant difficulties relating to internal assessment, such as ensuring reliability and authenticity, and preventing a too heavy workload. Marking reliability is a particularly important issue. When internal assessment makes a contribution to a high-stakes assessment system, it can place the classroom teacher in a difficult position, as both supporter and judge of a student's learning. This potential conflict of interest is compounded by a strong element of subjectivity surrounding the nature of the personal relationship between the teacher

and student. A teacher's judgment can be affected by past experience of a student's work, which establishes certain expectations. Teachers may sometimes be unclear about the limits of their role in guiding and supporting students 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 students' work, teachers may be heavily influenced by the general standards existing within their own schools.

For these reasons, moderators often have to make adjustments to teachers' internal assessment marks, even though the moderators may not have as comprehensive a view of student achievement as the teachers. Although the IBO provides support materials to teachers in how to guide and mark student work, it is not in a position to select and retain only those teachers able to mark consistently to the correct standard, as it can with external examiners. Despite such reservations, research has shown that internal assessment carried out by some teachers can be as reliable as external assessment (Black, 1993b).

Authenticity is another problem that raises questions over the reliability of internal assessment. Because of fears that people other than the student may contribute significantly to the work carried out, some assessment systems have either excluded internal assessment completely, or required that internally assessed tasks are carried out only under supervised classroom conditions. The IBO's view is that this represents an overreaction. For internally supervised but externally marked work, both the teacher and the student are required to sign a declaration of authenticity. Teachers must also sign a declaration that internally assessed work is the student's own. If evidence is subsequently found that the work is not genuinely the student's, then a judgment of malpractice becomes a possibility (IBO, 2003b). Plagiarism, particularly via the Internet, is obviously a major concern, and strong measures are taken to discourage, identify and penalize plagiarized work.

A third major concern over internal assessment is that of workload, particularly for the teacher and student. Internally set tasks are usually substantial and require a significant time commitment from the student. While it is appropriate for teachers to spend a considerable amount of time preparing students in the skills and processes required for internal assessment, there may be a strong temptation, felt by both student and teacher, to rehearse and practise the particular task set for internal assessment more than necessary, to make it as good as possible. The danger in doing this is that inadequate time is left for teaching and learning the remainder of the course, just as might be the case if too much class time is devoted to one particular syllabus topic. The perception may also arise in such circumstances that internal assessment is burdensome and too demanding. It is important for students to learn how best to manage their time and plan their own learning; this is, in fact, one of the broader aims of the DP as a whole, part of becoming an active learner.

Despite the acknowledged limitations in relying on internal assessment as the principal means of assessing student achievement, it should be noted that there are a number of national and state education systems that use only internal assessment, subject to varying degrees of external regulation, for end of secondary school, university entrance level, assessment. This is a reflection of the high value often attached to this mode of assessment.

To summarize, despite the many positive benefits of including internal assessment in a formal assessment system, there are good reasons for restricting the maximum contribution it should make to the overall DP assessment system, as is stated in the assessment policy. With a few exceptions, such as the arts subjects with an obvious highly practical slant, internal assessment forms a minor part of the assessment model for each subject.

4.3 Personnel

Each DP examination session is entirely dependent upon the considerable efforts of large teams of people. There are many thousands of teachers who prepare the students and contribute directly through internal assessment. There are about 4,000 examiners, who are distributed just as widely around the world as the teachers. Leading the examiners for each subject is a group of senior examiners, one chief examiner and one or more deputy chief examiners, depending on the size of entry and the assessment requirements for that subject.[1] Chief and deputy chief examiners normally each serve for a term of five years.

It is the duty of the group of senior examiners to prepare the set of examination papers and markschemes for each examination session, to lead the marking teams, to participate in the grade award meeting and deal with making judgments about results as part of the enquiry upon results service (see section 5.9). Chief examiners are normally appointed from the higher education sector, because of the subject expertise they can bring, because they know the requirements for university level study, and because there can be no conflict of interest in their role of judging student work. Deputy chief examiners are generally experienced examiners who are also practising DP teachers. They bring a considerable amount of practical experience to the assessment process.

Deputy chief examiners, and other examiners who may be called upon occasionally to help in examination paper preparation, are not allowed to work on preparing examination papers for the session for which their schools enter candidates. Thus, a deputy who is a teacher in a school entering candidates each May can help to prepare only November session papers. The chief examiner, together with other senior examiners and external consultants not linked to authorized DP schools, and IBO staff, have the necessary overview of the examination papers for both sessions in a year to ensure comparability. There are a few subjects, however, that do not have examination papers within the assessment model—for example, visual arts, theatre arts and TOK. Conflict of interest is less of an issue in these cases. All senior examiners mark student work for both sessions each year, but are not permitted to mark work from any school with which they have a connection, which may be personal, professional or geographical.

The chief examiners for each subject together form the examining board, which has an elected chair. The chair of the examining board chairs the diploma review committee (see section 4.2) and the final award committee (see section 5.7), as well as being a participant in other IBO committees, notably the Council of Foundation. One of these committees is the Diploma Programme committee, which has responsibility for overseeing the structure, regulation and implementation of the DP. The chief examiners for each subject group of the DP hexagon elect a group representative who sits on the diploma review committee. The other members of this committee are IBO staff and schools representatives.

Assistant examiners are appointed on an annual basis, though many of them are offered annual reappointments over a prolonged period. They are recruited largely from DP teachers around the world, but also from other educationists with suitable assessment experience in the appropriate subject. The activity of marking represents a significant opportunity for professional development for DP teachers, giving insights into how teachers in other schools prepare their students, as well as a much greater personal familiarity with the expected standards of work. In order to avoid possible conflicts of interest, assistant examiners are never given work to mark from schools with which they have any personal or professional connection.

The senior examiners work very closely with subject area managers/curriculum area managers, the IBO staff who are responsible for managing both the curriculum review and assessment in their subjects. Each subject area manager (SAM)/curriculum area manager (CAM) will be responsible for managing a small number of subjects. Curriculum and

[1] Smaller entry subjects, generally languages, may have only one examiner who does everything from setting the examination papers to marking and determining grade boundaries; this examiner is called the "examiner responsible" for that subject.

assessment activity for the DP is organized through the IB Curriculum and Assessment Centre (IBCA) in Cardiff, Wales, which is the largest of the organization's offices around the world.

The Cardiff office also houses the examination paper production department (EPPD), which has responsibility for taking the final drafts of examination papers, typesetting and formatting them and overseeing their translation, printing and shipment to schools. There is an examinations administration department (EAD), which is responsible for allocating student work to examiners, and making sure student work is marked on time and returned to the Cardiff office. EAD staff spend a lot of time working with DP coordinators (the member of staff in each school responsible for organizing examinations work) on matters such as student registration and administration queries. They also contact examiners about marking progress to ensure that all marking data is available for grade award meetings and all processing of marks is completed for the release of results. Finally, technical matters of assessment, such as moderation procedures and research, are dealt with by assessment staff under the assessment director, who has overall responsibility for all aspects of the formal DP assessment system.

5. Diploma Programme assessment processes

5.1 Examination paper preparation

The preparation of examination papers that have a high level of construct validity (that is, they appropriately represent the required range of skills defined by the course objectives over a balanced selection of course content), are at an appropriate level of intellectual demand, and are as free as possible from bias, is a challenging task. To meet this challenge, the examination papers, which are prepared anew for each examination session, are developed by a team of senior examiners and IBO staff, with input from external consultants, over a prolonged period. Work may begin on preparing a set of examination papers for a given course, 18 months to two years before the examinations are taken. The main stages of the process are described below for a typical large entry subject, with a team of senior examiners. For some small entry subjects, a single "examiner responsible" may carry out all the stages.

The first stage is the commissioning of the papers. Each member of the senior examining team (chief and deputy chief examiners, and sometimes other experienced examiners) is given the responsibility of preparing one or more papers to make up the complete set required for both higher level (HL) and standard level (SL) courses for a subject. The senior examiner may write the whole paper, or may compile the paper from questions submitted by other examiners. To maintain construct validity, there is a specification for each examination paper, detailing the number and type of questions (see section 3.5). This fixed structure also helps to ensure comparability of demand across successive versions of an examination paper. The senior examiners are always looking to set questions that have an appropriate variety of cognitive demand, and are also not too predictable. There should be a significant requirement within each examination paper for candidates to tackle questions and/or tasks that are in some way different from what they have done before (students become candidates when they have registered for the examinations). While there is a finite limit to the types of question that can be asked for any given course, the intent of at least some of them must be to require candidates to solve a problem or think creatively, to apply what they know in a new context rather than just to proceed with well-rehearsed skills or to restate knowledge.

Equal effort is put into the preparation of a markscheme to accompany each question paper. Even in the case of those papers containing open-ended tasks/questions, which are marked according to the same assessment criteria for each session (see section 5.4), explanatory marking notes are prepared that give guidance to examiners on how to apply the criteria in the context of each question. The markschemes are just as important to the integrity of the assessment process as the question papers. Each markscheme is much more than just a set of model answers; it provides guidance on how to mark common alternative approaches that candidates might adopt in answering a question, and also how to deal with commonly occurring errors or misconceptions that candidates might show.

After initial drafts of a set of papers have been prepared, the senior examining team and the relevant subject area manager (SAM) or curriculum area manager (CAM) will hold a paper editing meeting to review their work. At this meeting the set of papers for each level will be looked at as a whole, to see how course content and course objectives are covered across the suite of papers. The group will also critically review each question, looking at its appropriateness, correctness and any potential for ambiguity or bias. Details of the working of each question and its accompanying markscheme will be thoroughly scrutinized. Continuing discussions between the senior examining team about papers can take place in a secure electronic environment both before and after this meeting.

Once revised drafts have been agreed, the "post-meeting drafts" are sent to the external advisor, a consultant unconnected with the paper preparation process. The external advisor gives an independent view of the appropriateness of each set of papers in terms of course content coverage, level of difficulty in comparison to previous sets of papers, accessibility

within the time available for candidates, and all the factors previously considered at the paper editing meeting. The external advisor writes a report on the papers, which the SAM/CAM and senior examining team must consider. Decisions about how to respond to issues raised by the external advisor are made jointly by the paper author and the SAM/CAM, sometimes requiring the intervention of the chief examiner.

When the content of the examination papers and markschemes has been finally agreed, the examination paper production department (EPPD) takes over responsibility for scheduling and administering each set of examination papers through formatting, typesetting, printing and despatch to schools. The SAM/CAM and the senior examiner who wrote each paper will check, proofread and provide input at appropriate stages. For examination papers with a high level of technical content, an extra checker will read through the final proofs independently, answering all the questions and checking answers against the markscheme, to ensure that no errors involving technical terms, symbols or numbers have been introduced or overlooked. The extra checker must be an experienced examiner who has not been involved with the paper preparation process at all up to this stage, to avoid overfamiliarity with the content of examination papers interfering with detailed reading.

A breakdown of all the stages involved in the preparation of a typical examination paper is given in Table 1. Time must be allowed for the translation of examination papers in group 3–6 subjects into French and Spanish. This is normally only begun once the English version of the papers has been proofread by the examiner and nearly finalized, although issues raised by the translation process, particularly in regard to unanticipated ambiguity and subject-specific technical terms, can feed back into the English version of the papers.

Production stage	**Duration** (in weeks)
1. Papers commissioned and initial drafts prepared	14
2. Pre-meeting drafts submitted to EPPD and formatted	5
3. Paper editing meeting and revision	4
4. Post-meeting drafts submitted to EPPD, amendments made and checked	4
5. Post-meeting drafts to external advisor for report	5
6. Additional comments from SAM/CAM	3
7. Final drafts from examiners to EPPD/SAM/CAM	9
8. Final drafts checked by SAM/CAM	3
9. Final amendments, formatting and checking by SAM/CAM	4
10. EPPD proofread and cross-checking	3
11. Final proofread by examiner	4
12. Final amendments and proofread by SAM/CAM	4
13. Extra checker review and possible amendments	7
14. Final house style and consistency check	1
15. Sign off by SAM/CAM and sending to press	3
16. Return of printer's proof for checking	1
17. Examination papers printed and bagged for despatch to schools	12
Total time	**86 weeks**

EPPD—examination paper production department
SAM—subject area manager
CAM—curriculum area manager

Table 1: *Typical production schedule for a group 4 or group 5 examination paper.*

The schedule outlined in Table 1 is constructed to take account of a number of significant factors. First, at many stages in the process there is a need for careful consideration, discussion and reflection if high quality examination papers are to be produced. Second, many of the key personnel involved (SAMs/CAMs and senior examiners) can have a variety of other commitments that prevent them from attending to examination paper preparation as soon as the materials are ready for the next stage. Third, at critical points in the process it is necessary for all of the papers in a set to be ready together before they go on to the next stage. Therefore, a set of papers may be held up by a delay on just one of them. Fourth, the quantity of papers to be prepared means that immediate attention cannot always be given by EPPD staff to each set. For a May session, there are over 700 different examination papers to prepare, including the translations. For November, there are approximately 400 papers. With two examination sessions per year and the extended preparation schedule, examination paper preparation department staff are likely to be making progress on up to 2,000 examination papers at any one time, meaning that each one cannot be given immediate attention as soon as it is ready for the next stage. Similarly, the large number of different papers, and the high volumes to be printed for some of them, mean that the printers need a considerable period in which to print and collate all the papers. The production schedule is aimed at ensuring the examination papers arrive in schools about three weeks before the start of the examination schedule. This is to allow for any delays that might occur in a global delivery network, particularly relating to customs clearance. It also gives time for schools to check their deliveries and for any discrepancies to be rectified.

Owing to resource implications and the difficulty of identifying an appropriate trial group, pre-testing of examination papers is not carried out.

5.2 The examinations

The examinations for each session take place over a period of approximately three weeks in May and November. Given the number of subjects to be examined, this can only be done if two or three different subjects are sometimes examined in the same slot in the schedule. The choice of subjects to be blocked together in this way is made so as to reduce to a minimum the number of clashes that are likely to occur for individual candidates. However, a small number of such clashes are inevitable and procedures for dealing with these are described in the *Vade Mecum* (the procedures manual for Diploma Programme [DP] coordinators and teachers).

Examinations are scheduled to avoid more than six hours of examining in a single day where possible, under normal circumstances. Friday afternoons are kept free from examinations as a means of extending at least some consideration to those schools whose working week is not Monday to Friday. The normal pattern for the examinations relating to a particular course is to schedule the two or three papers consecutively, starting one afternoon and finishing the next morning. This arrangement is preferable to presenting all the examinations for a given course on the same day for the following reasons.

- For most candidates this spreads the examinations more evenly over the three-week schedule. It allows candidates the opportunity to recover if they feel they have not done themselves justice on a particular occasion.
- If a candidate is ill or unavoidably absent on a given day, there is still a possibility that the remaining paper(s) can be taken on the day before/after. If sufficient assessment components have been taken, it may still be possible to award a result to the candidate in some circumstances.
- The different answer papers (scripts) will be sent to examiners on different days, reducing the possibility of all the scripts being lost in transit.

To complete the schedule over three weeks, without using Friday afternoons, not all examinations can follow the above sequence. Language A1 and language A2 examinations are sometimes held so that paper 1 and paper 2 are separated by a period of days. Since these papers are quite independent of each other in terms of content, and the language involved is one in which the

candidate is highly fluent, it is considered that the separation of papers will have less of an impact than in other subjects.

Schools must conduct examinations according to a strict set of regulations laid out in the *Vade Mecum*. These regulations cover everything from dealing with receipt of examination material, through holding the examinations, to sending the scripts off to examiners. Random inspection visits are paid to schools by IBO regional office staff and consultants during the examination schedule, to check school administrative procedures and security arrangements.

The *Vade Mecum*, together with a publication addressing the concerns of candidates with special assessment needs, provides information to schools on how to put in place special assessment arrangements for candidates with individual needs, such as a specific learning difficulty, a behavioural difficulty, a physical, sensory or medical condition, or a mental health problem.

5.3 Internal assessment and other non-examination components

Internal assessment can take a variety of forms, from an individual oral presentation and discussion lasting ten minutes for language B courses, to a research workbook in visual arts which is each student's personal record of their artistic development, recommended to require 72 hours of work at higher level (nearly one-third of the course). In between are cases such as the experimental sciences (group 4) internal assessment, made up at higher level of pieces of work selected from a portfolio of 60 hours of practical work and investigations (25% of total teaching time). The nature of the assessment task reflects the purpose of the internal assessment, in particular the emphasis on and type of process skills involved. This is especially the case in group 4, where particular pieces of practical work that meet certain criteria should be selected from the whole portfolio.

However, there are certain procedural features that are common to all internal assessment. First, internal assessment should, as far as possible, be woven into normal classroom teaching. Internal assessment focuses on skills, not subject content, but the internal assessment activities chosen by the teacher or the student can often be used as vehicles for teaching prescribed course content. Activities to be used for internal assessment can also be used to develop skills, that is, formatively, as well as to contribute summatively to the final assessment outcome. The decision about when to make the transition from formative assessment to using an activity as part of final summative assessment is often left up to the teacher. Internal assessment should not be viewed as a separate "bolt-on" activity to be conducted after a course has been taught.

A second common feature is the prescribed level of support given by the teacher to the student for activities that will actually contribute to final assessment. Where the end result of the activity is a relatively formal piece of written work, teachers are generally permitted to discuss the topic and approach with the student and give restricted advice on a first draft. Any subsequent amendment or editing must be by the student, so that the final work submitted for internal assessment is the student's own. Sometimes, group activities are permitted as a basis for internally assessed work, but where written work is to be submitted this must always be the individual work of each student.

Third, internal assessment is conducted by applying a fixed set of assessment criteria for each course (see section 5.4). These criteria describe the kinds and levels of skills that must be addressed in the internal assessment. Teachers should ensure that students are familiar with the internal assessment criteria and that the pieces of work chosen for use in internal assessment address these criteria effectively. This is especially significant in the group 4 experimental sciences, where the portfolio of practical work might quite legitimately contain a number of pieces of work that are not suitable for use against the assessment criteria. The group 4 internal assessment criteria are intended to address a particular set of skills that may not be evident in some standard science laboratory work.

The last two of these features also apply to the small number of externally marked, non-examination components. These include extended essays, theory of knowledge essays, language A1 world literature assignments, language A2 written tasks and music investigations. Although these pieces

of work are sent to examiners for marking and are not marked by teachers, the teacher's role in discussing the work with the student, lending advice and considering the assessment criteria, is very similar to that for internal assessment components, which are marked by the teacher.

5.4 Marking

Mention has previously been made of the importance of marker reliability, that is the ability of an assessment process to provide almost the same mark to a piece of work, regardless of which examiner marked it and on which occasion it was marked (see appendix A.3). There are three main ways of ensuring this. First, it is important to appoint and retain only those examiners who can mark consistently and objectively. The great majority of DP examiners are experienced teachers of the programme. Such examiners are ideally suited to the task, through having prior familiarity with the course being taught and its assessment requirements, and some knowledge of the expected standards. Second, all examiners, except the most senior one for each component, have their marking checked every examination session—past good performance does not guarantee what may happen in the next session. This procedure is called moderation and is described in section 5.5. The third method considered in more detail here, is to provide examiners with comprehensive instruction on how to go about marking. This can be done through prior training, an area of activity that is planned for much greater development by the IBO in the near future, through electronic means. Diploma Programme examiners receive detailed instruction about the administrative procedures to be followed that will allow successful moderation to take place, and also substantial information about how to allocate marks.

The IBO uses two principal methods of guiding examiners in the allocation of marks: analytic markschemes and assessment criteria (which itself has a variant called markbands).

5.4.1 Analytic markschemes

Analytic markschemes are prepared for those examination questions that expect a particular kind of response and/or a given final answer from the candidates. These markschemes give specific instruction to examiners regarding how to break down the total mark available for a question for different parts of the response, which reflects the importance given by the senior examining team to those different parts of the response. Candidates may get different parts of a question right or wrong and lengthy structured questions are designed 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 him/her. This is the main reason for using structured questions in technical subjects such as science and mathematics, to enable 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 and proceeded in quite the wrong direction. The most highly elaborated analytic markschemes are found in mathematics. Such markschemes contain specific instructions on how to mark particular kinds of incorrect answer, and how to deal with following through candidates' working when they have made a mistake on part of a question.

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. Examination papers will always contain at least some questions where 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.

As well as being provided with markschemes and assessment criteria, assistant examiners are expected to receive advice from senior examiners, by telephone and/or e-mail, during the marking period itself.

At the time of writing an examination paper and its accompanying markscheme, the senior examining team may not always be able to anticipate all the common kinds of response that candidates will provide. Although every effort is made to predict the likely range of responses,

this is particularly difficult on a global basis given the variety of educational cultures and teaching styles that exist around the world. To address this problem and reduce as much as possible the dependence on possibly variable examiner judgment, the senior markers for each examination paper in large entry subjects will meet soon after that examination has been taken and review the scripts of a selection of candidates. This is called a standardization meeting, its purpose being to make a small number of final additions and amendments to the markscheme in the light of actual responses written by a number of candidates, and to ensure that the senior markers have an agreed interpretation as to how the markscheme should be applied. This is crucial to successful moderation, and issues arising from this meeting are communicated directly to all assistant examiners.

5.4.2 Assessment criteria

Where an assessment task is so open-ended that the prospective variety of valid responses is too great to permit analytical markschemes to be written, then assessment criteria are applied instead. Assessment criteria do not refer to the specific content of a candidate's answer, although some may refer to the need for candidates to show specific kinds of content knowledge. The criteria concentrate more on the generic skills that candidates are expected to demonstrate, regardless of the specific individuality of the response. For example, the five assessment criteria for the language A1 examination requiring a written commentary on one of two unseen text extracts are titled: understanding of the text, interpretation of the text, appreciation of literary features, presentation and formal use of language. Each criterion comprises a set of related skills that candidates are expected to demonstrate at a range of levels of accomplishment. In language B (the main second-language course), on the examination paper requiring written production in response to a choice of tasks, the three assessment criteria are: language fluency and accuracy, cultural interaction (that is appropriateness of style, register, devices and structure in relation to the intended audience) and communication of message. As a third example, the four criteria for philosophy essay questions based on optional themes of study are: clarity of expression, knowledge and understanding of the philosophical issues, identification and analysis of relevant material, and development and evaluation.

Because of their highly variable nature, internal assessments and externally assessed non-examination tasks are also marked using assessment criteria. For the internal assessment of all group 4 experimental sciences, there are eight criteria:

- initial planning of investigation (defining the problem, forming a hypothesis and selecting variables)
- selecting apparatus and designing a method
- data collection
- data processing and presentation
- conclusion and evaluation
- manipulative skills (carrying out techniques and following instructions)
- team-working skills
- motivation and ethical working.

There is a close relationship between these criteria and the course objectives given in section 3.5, supporting high construct validity (see appendix A.1).

In all cases where assessment criteria are applied, differences in candidate achievement that lead to the award of different marks are defined by achievement level descriptors for each criterion, which describe the typical ways in which a candidate's response can be measured against the criterion. The total possible mark for a piece of work is arrived at by adding together the maximum achievement level for each criterion. Greater weighting is given to criteria considered to be more important by giving them a greater number of achievement levels.

It is important to note that, although criterion level descriptors are hierarchical in nature, and often deal with the hierarchy of cognitive skills defined by Bloom *et al* (1956), the two hierarchies are independent. 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. It is recognized that there is a range of levels of achievement within each of the cognitive skill areas.

The following achievement levels for the criterion "interpretation of the text" for the language A1 written commentary paper (IBO, 1999, p44) serve as an example. This criterion measures the relevance of the candidate's ideas about the text, how well the candidate has explored these ideas, how well the candidate has illustrated his/her claims, and the extent to which a candidate has expressed a relevant personal response. The achievement levels illustrate the very close link between the first objective given for language A1 in section 3.5 and the assessment process conducted at the end of the course. The objective is "…candidates will be expected to demonstrate an ability to engage in independent literary criticism in a manner which reveals a personal response to literature."

Achievement Level

0 **The candidate has not reached level 1.**

1 **Little interpretation of the text**

- the candidate's ideas are mainly insignificant and/or irrelevant **or**
- the commentary consists mainly of narration and/or repetition of content.

2 **Some interpretation of the text**

- the candidate's ideas are sometimes irrelevant
- the commentary consists mainly of unsubstantiated generalizations **or**
- the commentary is mainly a paraphrase of the text.

3 **Adequate interpretation of the text**

- the candidate's ideas are generally relevant
- the analysis is adequate and appropriately illustrated by some relevant examples.

4 **Good interpretation of the text**

- the candidate's ideas are clearly relevant and include an appropriate personal response
- the analysis is generally detailed and well illustrated by relevant examples.

5 **Excellent interpretation of the text**

- the candidate's ideas are convincing and include an appropriate and considered personal response
- the analysis is consistently detailed and persuasively illustrated by carefully chosen examples.

The approach used in DP assessment in the application of criterion achievement levels is a "best fit" model. The examiner or teacher 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, in a way that the maximum mark on an analytic markscheme probably would (analytic markschemes operate over a much greater mark range than do assessment criteria).

A number of examination tasks are marked according to the same assessment criteria each examination session. Although the general nature of the task (usually an essay or piece of extended writing) remains the same in each examination session, the specific requirements of each question may have implications for the way in which the assessment criteria should be applied. Analytic markschemes are not appropriate in such cases, but marking notes are generally written by the senior examiner who prepared the examination paper. These marking notes give guidance to assistant examiners on how the assessment criteria should be applied to each question. 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.

5.4.3 Markbands

It is sometimes not judged appropriate to separate out the different assessment criteria that can be applied to a particular piece of work. Assessment criteria are most appropriately applied in relative independence from each other, with candidate performance on one criterion not influencing performance on the others. In practice, this can rarely be fully achieved, but if a situation arises where it is not possible to discern separable assessment criteria, then a different approach is adopted. This may also be necessary where the work to be assessed is so variable that a set of criteria, each of which is readily applicable to all responses, cannot be derived. 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 number of marks.

The descriptors themselves tend to be fairly lengthy, covering a range of potential qualities evident in candidates' work, and will again relate directly back to the course objectives. Examples of markbands can be found in the Diploma Programme *History* guide (IBO, 2001a). As with assessment criteria, a "best fit" approach is used, with markers additionally making 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, Ch 5).

5.4.4 Marking schedule

The external marking carried out by examiners takes place within demanding time constraints. There are six weeks between the date of the last examination and the date of results release. In the period following the day an examination is taken, the markscheme may have to be finalized, candidate scripts must reach the examiners who may well be in different parts of the world, the scripts must be marked and then sent back to IBCA, a sample of each examiner's marked scripts must be sent to a more experienced examiner who moderates them and sends them to IBCA, the grade award meeting must be held, and any re-marking of scripts found to be necessary must be completed. The moderation of each examiner's marking is based on a sample of their earlier work. This sample is moderated while the examiner continues to mark the remainder of his/her allocation of scripts. The moderation process itself is described in the following section, but it should be noted that additional methods are in place to check the consistency of examiners' marking throughout their whole allocation after they have sent off their moderation sample.

First, at the grade award meeting, the senior examiner team will look at a wide range of scripts when determining grade boundaries. While the emphasis of this process is to consider candidates' work, rather than the marking, cases of aberrant marking will become obvious. Second, after the grade boundaries have been established, a selection of candidates who have just missed gaining a higher grade will have their work re-marked by senior examiners. Again,

unsatisfactory original marking can be detected at this stage. Third, during the re-marking of candidate work known as "enquiries upon results", (a service that can be requested by schools after results have been issued), evidence may again emerge that indicates unsatisfactory and inconsistent marking by an examiner. There are, therefore, a number of additional means of checking the quality of an examiner's marking apart from the moderation sample.

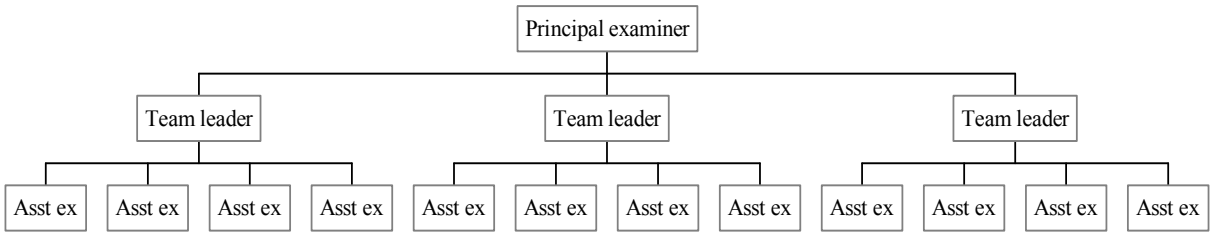
5.4.5 Visiting examiners

Finally, reference has been made throughout this section to the remote marking by examiners of assessment materials, generally written scripts, but also on occasion videotapes and audio cassette tapes. However, there are certain circumstances in which examiners visit the school and assess candidate work directly. Because of high resource costs and the difficulties of maintaining inter-examiner reliability when examiners are widely and thinly distributed around the world, visiting examiners are only used where there is a clear and unequivocal requirement for them. Currently, visiting examiners are used only for visual arts and a pilot dance course. The requirement to see candidates' work *in situ* and to discuss their work directly with them overrides the limiting constraints. Visiting examiners make their judgments according to assessment criteria, and their judgments are moderated by means of photographs or videos. This represents a significant compromise between the needs for reliable assessment and for valid assessment of candidate achievement. The procedures for dealing with marking by visiting examiners are under constant review.

5.5 Moderation

5.5.1 Moderation procedure

Moderation is the principal tool for ensuring marking reliability, though not the only one (see section 5.4.4). Every examiner, except the principal examiner for each component who sets the standard, sends a sample of their marking to an experienced and reliable examiner, who acts as team leader. This sample is re-marked by the team leader and a statistical comparison of the paired set of marks determines whether the original examiner's marking is acceptable, perhaps with some slight adjustment, or unacceptable. Moderation is a hierarchical process and the different levels of the hierarchy for a typical large entry examination component are illustrated in Figure 2.



Asst ex—assistant examiner

Figure 2: *Moderation hierarchy for a typical large entry, externally assessed component. The diagram has been simplified for ease of representation—most teams actually consist of about 10 assistant examiners.*

The principal examiner for a component is often the chief examiner or deputy chief examiner, but can also be a highly experienced and reliable former team leader. Generally, the principal examiner is also the author of the examination paper or was greatly involved in setting that paper. Team leaders are experienced examiners who have shown that they can mark consistently and accurately over a number of examination sessions. Each team leader will generally oversee up to 10 assistant examiners. For examination paper components, the moderation sample size is 15% of each examiner's total allocation of scripts, with a minimum of 10 and a maximum of 20 scripts. Examiners are requested to submit samples covering a number of schools and as full a range of marks as possible.

5.5.2 Correlation criterion

The pairs of marks relating to each candidate's script in the sample are subject to statistical analysis. One statistical measure is the correlation coefficient (the product moment correlation coefficient is used). This measures the consistency of the relationship between the two examiners' marking. A correlation coefficient of zero indicates no relationship at all; a score of one indicates perfect consistency in the relationship between the two examiners' marking and agreement in ranking candidates from best to worst (though not necessarily requiring the two examiners to give exactly the same marks). A coefficient of –1 indicates consistently opposing views between the two examiners with regard to the relative merits of candidates' work, the examiners producing opposite rankings to each other.

For an examiner's marking to be acceptable, the correlation coefficient must be at least 0.90, indicating a high level of agreement between the assistant examiner and team leader. If the correlation coefficient is less than 0.90, it is most likely that the assistant examiner's allocation of scripts will be re-marked by a more reliable examiner. The assistant examiner will not be used again in such circumstances, unless a particular cause can be identified that gave rise to the lack of agreement and can be readily remedied.

However, a high correlation coefficient on its own is insufficient to make an examiner's marking acceptable. It is possible that an assistant examiner consistently awards, for example, four marks too many to every candidate, on an examination marked out of 25. This would generate a very high ranking correlation, but does not indicate satisfactory marking.

5.5.3 Linear regression

A further analysis is carried out on the data for each moderation sample, which permits an appropriate average adjustment to be applied to all of an examiner'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 assistant examiner and the team leader. This is illustrated in Figure 3.

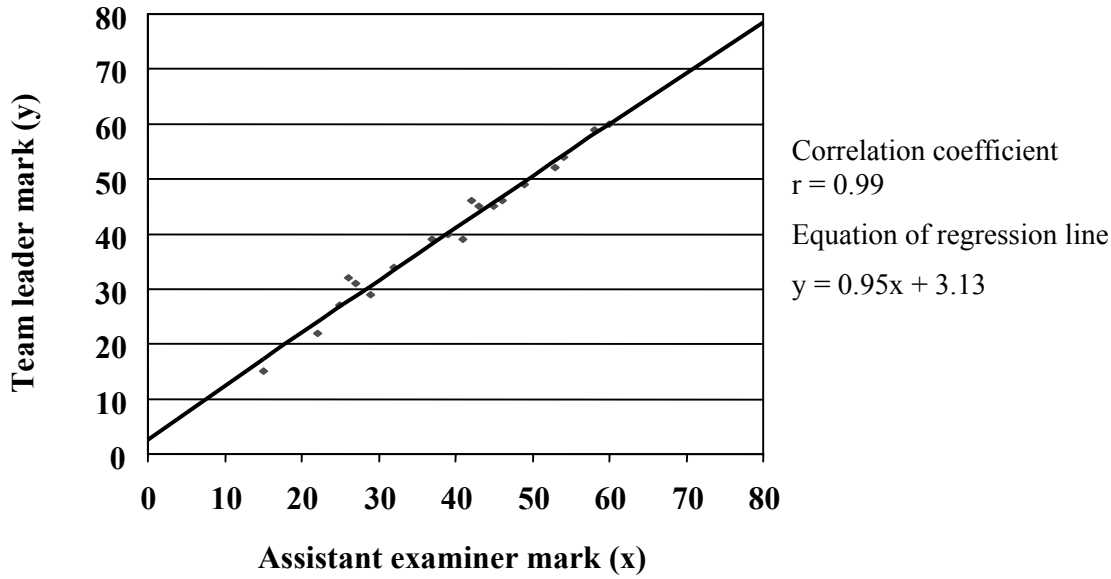


Figure 3: *The moderation regression line for an assistant examiner on a paper marked out of 80. Each individual point represents the pair of marks given to a sample script by the assistant examiner and the team leader. The continuous regression line is used to convert the assistant examiner's marks into moderated marks.*

The equation of the regression line calculated from the sample data can be used to convert each mark (x) awarded by the assistant examiner into an equivalent mark (y) that the team leader 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, must inevitably be derived from general trends apparent in the marking. Individual variation relating to particular candidates cannot be accounted for. For example, one of the individual points in Figure 3 shows that the team leader gave a mark of 46 to a script that the assistant examiner gave 42. However, such an adjustment is different from the general trend and would probably not be appropriate for every candidate that the assistant examiner gave 42. Instead, the regression line, reflecting the average trend of marking difference, would moderate each mark of 42 given by the assistant examiner to 43. It is for this reason that further checks on marking, particularly "at risking" (see section 5.6), are undertaken. 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.

5.5.4 Other criteria

As well as achieving a satisfactory correlation coefficient, each examiner's marking must also satisfy two other criteria before a moderation adjustment can be applied automatically. The slope of the regression line must be between 0.5 and 1.5. If the slope of the line is too low (or too shallow), it means that the assistant examiner 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 team leader has had to compress the assistant examiner's mark range considerably. If the slope is greater than 1.5, the line is too steep and the opposite applies; the assistant examiner has not differentiated sufficiently between poor and good candidate work and the team leader has had to expand the mark range awarded.

The second criterion is that the difference between the mean assistant examiner sample mark and the mean team leader sample mark must be less than 10% of the total mark available for that component. Thus, if the total mark available for a component is 30, the mean assistant examiner mark must be within three marks of the mean team leader mark for the given sample of work. An examiner whose marking fails to meet either of these two criteria may be consistent but clearly is out of line with the expected standards of marking and so is treated as a moderation failure.

5.5.5 Moderation failure

All cases of examiners who fail moderation are reviewed individually by assessment staff at IBCA, who consider the underlying data carefully and may decide:

1. to apply some other kind of moderation adjustment, perhaps different linear adjustments for different parts of the mark range
2. to request further sample data in order to clarify the trend
3. to require a complete or partial re-mark of that examiner's work.

In all such cases, a recommendation will be made by the SAM/CAM on whether to retain the services of the examiner for future sessions.

5.5.6 Compound moderation

It should be noted that team leaders also have to submit a sample of their marking, and may themselves be subject to a moderation adjustment through linear regression. The moderation equations derived for the different levels of the hierarchy illustrated in Figure 2 are compounded to give an overall moderation adjustment applied to each assistant examiner. Such compounding works satisfactorily, provided the upper levels of the moderation hierarchy are in close agreement with each other—that is, that team leaders are marking to a standard very close to that of the principal examiner.

5.5.7 Adaption of linear model

The assumption behind the present moderation adjustment system that a straight line is the best model for adjustment is open to question. The IBO is currently considering a best-fit curve, rather than a best-fit straight line. Comprehensive testing to ensure that inappropriate curved functions are not derived and that such curved functions can be combined satisfactorily across levels of moderation is needed before such a change can be made. In the meantime, the straight-line model 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 maximum marks or zero (see, for example, Figure 3). A candidate whose work is poor but worth a few marks may be given zero through moderation, or as would be the case from Figure 3, a candidate whose work is genuinely worth zero might be given a few marks when nothing of any worth had been written on the script.

Similarly, at the other extreme, a candidate whose work is very good but contains a few clear failings, may be given full marks, or a candidate who genuinely deserves full marks may not be given them. The impact of a moderation adjustment based on a straight line is generally greatest at the extremes of the mark range, while there is most confidence about moderation adjustments in the middle of the mark range, where most of the sample data tends to be concentrated.

To overcome this problem, "tailing" is applied to marks in the top 20% and bottom 20% of the available mark range. At these extremes, the calculated regression line is modified and substituted by new "tailed" lines that link from the regression line to the maximum coordinates and minimum coordinates, as shown in Figure 4.

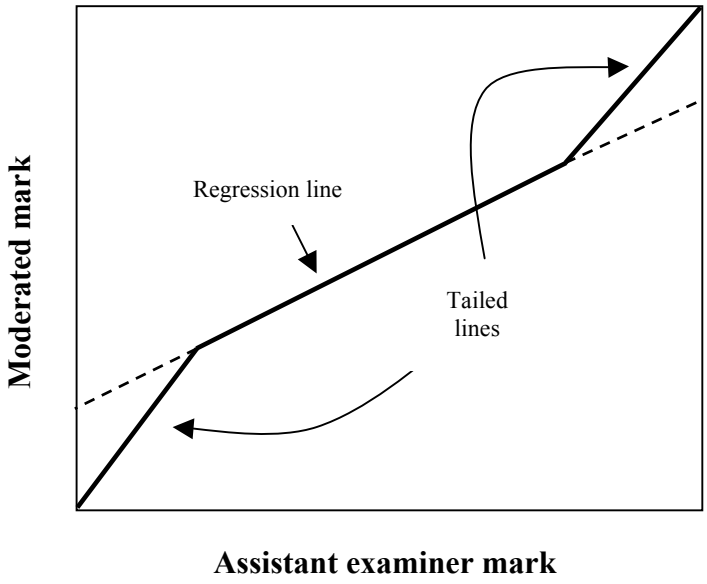


Figure 4: *"Tailing" of a regression line to prevent candidate marks from being adjusted away from or towards maximum and minimum values.*

The rationale for implementing this procedure is that, however generous or harsh an assistant examiner is, he/she cannot award marks below the minimum or above the maximum. A generous examiner, for example, could end up giving a maximum mark to candidates who do not quite deserve it, as well as to those who do. A moderation process must treat all candidates on the same mark in the same way, and it is preferable to give such candidates the benefit of the doubt when a spread of achievement has been compressed to a single mark.

In this way, candidates awarded the maximum by the assistant examiner (who are very few in number and may even deserve a mark exceeding the maximum), retain that maximum mark whatever the moderation adjustment, and no other candidates in that examiner's allocation can be awarded the maximum mark. This assumes that the assistant examiner's marking is good enough to pass through the automatic moderation process. If the examiner cannot be automatically moderated, tailing is not applied. At the other extreme, a moderated zero mark can only be derived from an original zero mark. Tailing 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.

5.5.8 Moderation of internal assessment

The moderation of internal assessment, for which the original marking is done by classroom teachers, operates according to a similar basic structure but with some differences in the criteria for passing or failing moderation. (Note that passing moderation does not mean that the original marking was absolutely accurate; it only means that it was consistent enough for linear moderation adjustments to be applicable.) The moderation hierarchy for internal assessment is represented in Figure 5.

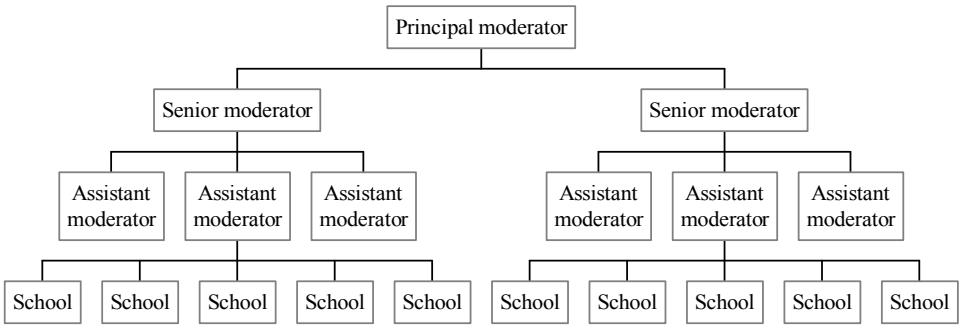


Figure 5: *The hierarchy of moderation applied to internally assessed components. The diagram is simplified. In reality, each assistant moderator is likely to receive internal assessment samples from about 10 schools, and each senior moderator is likely to oversee about 10 assistant moderators.*

The slightly different approach to the role of internal assessment moderation is indicated by the different titles given to the examiners who carry it out—that is, principal moderator, senior moderator and assistant moderator. The great majority of these are experienced DP teachers. All internally assessed components are marked by applying assessment criteria, 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 moderator can have. Because of this, moderators for most internal assessment components, except for language orals, 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.

Despite this more flexible approach, the great majority of all moderation failures arise from internal assessment. The IBO does not accredit teachers in regard to their competence in conducting internal assessment, and is not in a position to control which teachers conduct internal assessment in the same way that it can with assistant examiners. Also, it can be difficult for a teacher to submit a satisfactory moderation sample covering a wide range of marks or types of work, given the restricted and sometimes non-representative or very homogenous nature of student groups within a class at a given school. If the sample is unavoidably restricted, it becomes more difficult for that sample to meet the statistical requirements for processing through the automatic moderation system.

Because of such factors, and because internally assessed work is inherently more open-ended and more open to subjectivity in marking, the criteria for passing automatically through moderation are slightly less stringent. The correlation coefficient must be at least 0.85. The gradient of the regression line must still be between 0.5 and 1.5 as for external assessments, but there is no limit to the difference in sample means. Tailing of the regression line is retained at the lower end of the mark range, but not applied at the upper end because it more frequently happens that teachers have awarded maximum marks to candidates who clearly do not deserve them. Moderation sample sizes are smaller for internally assessed components (ten, eight, five, or fewer than five sent from the schools to assistant moderators, according to the number of candidates in the school group). This is partly because larger sample sizes would, in effect, result in the re-evaluation of most candidate work, and partly because of the smaller allocations of candidate work given to moderators and senior moderators. The internal assessment work for a single candidate is generally more substantial and more demanding to review than an examination script, meaning that allocations of work to moderators have to be smaller. The IBO selects the candidates whose work comprises the internal assessment moderation samples after the school has formally submitted its internal assessment marks.

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 IBO expects 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.

5.6 Grade awarding and aggregation

5.6.1 Procedure

The grade award meeting represents the culmination of the assessment process for each subject. Taking place about 35 days after the date of the examination papers, by which time all of the marking and moderation for the subject should be complete, grade award meetings require the senior examining team for each subject with a reasonably large candidate entry to convene at IBCA. The team, together with the SAM/CAM, reviews the operation of the assessment components for the session, sets the grade boundaries for each of the higher level and standard level courses, resolves any outstanding issues relating to examiner marking, and carries out “at risking” (see section 5.6.5).

The first task of the grade award meeting is to reflect on the operation of each component. All participants will have been actively involved in marking at least one component and most will have contributed to writing the examination papers. In addition to their own experience, the senior examiners will review the comments formally submitted by teachers about the examination papers and the reports from assistant examiners on the general nature of candidate responses. This information provides both an important background for the senior team to help them agree appropriate grade boundaries in the light of general perceptions about the examination papers, and also a basis for writing the subject report, which is distributed to all schools (see section 5.9).

Following this, the team will consider each component for which new boundaries must be set every session. The boundaries for internally assessed components, and externally marked non-examination components, are not revised each session. They are normally set only once, when such components have just been introduced or revised. On the other hand, new boundaries will be set for each examination paper at each session. The change in boundary marks is normally slight because every effort is made to construct each new version of an examination paper at about the same level of overall difficulty as its predecessor.

5.6.2 Setting grade boundaries

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 before gaining substantial experience of how candidates have actually responded. Sometimes, a question that was set to be quite demanding turns out to be readily answerable by candidates, or the reverse occurs. Pre-testing of examination questions might help to reduce this element of unpredictability, but final judgments about how difficult a question actually turns out to be would still normally have to await the actual outcome of the live marking.

The setting of grade boundaries is an extended matter requiring considerable deliberation and the reconciling of information from different sources: the experienced judgment of senior examiners, statistical comparisons and the expectations of experienced teachers, who are familiar with the standards and know the candidates better than anyone.

The most significant grade boundaries for each examination paper—those between grades 3 and 4, between grades 6 and 7, and between grades 2 and 3—are determined judgmentally in that order. These are the boundaries that have the greatest impact on candidates' progression into higher education. The remaining boundaries are determined by interpolation from these judgmentally set boundaries. The principal means of setting judgmentally determined grade boundaries is by review of the quality of candidate work against grade descriptors, which are generic descriptions of the standard of work expected of typical candidates at each grade. Grade descriptors attempt to characterize the typical performance of candidates expected for a given grade and are intended to give some guidance to classroom teachers on how to prepare their students and how to make predictions of their likely grades. However, it must be accepted that such descriptors can only be interpreted fully in the light of much experience in their application to actual pieces of assessment work. An example of a grade descriptor, for grade 4 in mathematics courses, is as follows:

> **Grade 4 Satisfactory performance**
>
> Demonstrates a satisfactory knowledge of the syllabus; applies mathematical principles in performing some routine tasks; successfully carries out mathematical processes in straightforward contexts; shows some ability to recognize patterns and structures; uses problem-solving techniques in routine situations; has limited understanding of the significance of results and attempts to draw some conclusions; communicates mathematics adequately, using some appropriate techniques, notation and terminology; uses technology satisfactorily.

Before the grade award meeting the senior examining team, and in particular the principal examiners, 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 the meeting about how each paper has functioned and a brief consideration of the moderated mark distribution (histogram) for the component in comparison to previous years, lead to the team considering a range of marks within which the grade boundary is thought to lie. A number of scripts at each mark will be considered in turn until the meeting agrees that, for example, 27 marks represent the highest mark for grade 3, and 28 marks the lowest mark for grade 4.

During this process, the senior examiners are asked to focus, not on the marks present on the script, but on the nature of actual candidate responses and how well these match the grade descriptors. For this review, scripts are selected as much as possible 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. Once the meeting has agreed on a boundary, the process is repeated for the remaining boundaries and the other components. The members of the meeting have access to scripts from previous examination sessions that lie exactly on the grade boundaries, to help them maintain a consistent standard from session to session.

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. For example, in group 4 experimental sciences, paper 1 is a multiple-choice question paper. Paper 2 is made up of different kinds of question, but based on the same course content as paper 1. Thus, if boundaries are determined from paper 2 that give, for example, 9% of candidates in grade 7, then boundaries are calculated for paper 1 that also give as closely as possible 9% of candidates grade 7. Thus, candidate performance on the multiple-choice paper has little effect on the overall grade distribution for the subject, but will clearly have an impact on the subject results of individual candidates.

5.6.3 Aggregation

When the grade boundaries have been established for all components, provisional overall subject grade boundaries are calculated and a provisional subject grade distribution inspected. In order to aggregate the marks (and boundaries) from the different components to form a percentage total mark, they may first 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. For example, a higher level course in a subject may be made up of three components and the model requires that component 1 contributes 50% to the final result, component 2, 30% and component 3, 20%. If component 2 is designed to have a total available mark of 90, then these marks, after moderation, would have to be scaled by dividing by three to achieve the required weighting of 30%. The same applies to the grade boundaries set for the component, which also would have to be determined initially out of 90 marks.

A significant factor here is that the concept of weighting refers to possible achievement (represented by available marks) and not distribution or spread of marks awarded. A component weighting of 30% does not mean that this component must contribute 30% to the final spread of candidate marks—such an approach is more typical of norm-referenced systems, which focus on differentiation between candidates as the primary goal of each assessment item and test. In the DP assessment system, candidate differentiation is a secondary consideration to candidate achievement. A component that does not spread candidates out in terms of marks gained, but still records significant educational achievement, makes a valuable contribution to the overall judgment of candidate performance. Different components of the DP assessment model for a subject level may produce different spreads of candidate marks. It is possible for different components to have quite different grade distributions (the percentage of candidates in each grade).

It is also not a requirement that a component with a given weighting should contribute exactly that weighting to candidates' total marks. The significance of the weighting is to indicate the proportion of overall credit available for a given component, not necessarily the overall proportion of achievement.

The approach adopted in the DP assessment system may not reflect the more sophisticated methods of weighting, combining (aggregating) and scaling described by, for example, Wood (1991, Ch 10), but is based on the sound criterion-related principle that no candidate's mark should be modified solely on the basis of how other candidates around him/her have fared. The approach also has the considerable benefit of transparency in comparison to more complex models of weighting and scaling.

After applying any necessary scaling, the marks and boundaries are rounded to the nearest whole number. Marks or boundaries ending in exactly 0.5 are rounded in the candidates' favour—a mark is rounded up, while a boundary is rounded down. Then, the highest mark in each component grade is added together to give the highest mark in the corresponding subject

grade. The highest mark in grade is chosen in preference to the lowest mark in grade to counteract the so-called "regression to the mean" effect. In simple terms, the regression to the mean effect states that it is more difficult to achieve a certain grade across a number of components than it is to achieve it in a single component. For example, it would be possible to have an assessment model with three components, in which 10% of the candidates gained a grade 7 in each component, while fewer than 5% gained a grade 7 overall, depending on how consistent candidates' performance was across the three components.

An example of the aggregation is given in Table 2. From the data given in the table, the overall subject 6/7 grade boundary would be 82/83. Thus, a candidate gaining the highest mark for grade 6 in two components, and the lowest mark for grade 7 on the third, would be awarded a grade 7, and so on. 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.

Grade boundary	Component 1 (50%)	Component 2 (30%)	Component 3 (20%)	Overall subject boundary mark
7	43	25	17	
6	42	24	16	82
6	38	21	15	
5	37	20	14	71
5	31	18	12	
4	30	17	11	58
etc	etc	etc	etc	etc

Table 2: *An example of the aggregation of scaled boundaries in the IB Diploma Programme.*

A principle of compensation across the components is followed within each DP subject. This means that a poor mark gained by a candidate on one component can be offset by a higher mark on another. There are no "hurdles" that have to be met on given components to achieve particular subject grades, apart from a requirement that a candidate must submit work for each component to be awarded a subject grade. Thus, it may be theoretically possible, though highly unlikely, for a candidate to gain zero marks on one component, but still achieve a subject grade 7 if the marks are high enough on the other components. (Note, however, that the same principle of compensation does not apply when subject results are combined to determine whether a candidate should be awarded the diploma. In this case, a system of hurdles relating to minimum grades in each subject, in TOK and in the extended essay, is applied.)

5.6.4 Grade distribution

After the aggregation of component marks and grade boundaries by computer processing, the grade award meeting will review the provisional subject grade distribution before confirming the final grade boundaries. Comparisons must be made with previous years' results, with the distribution of predicted grades received from schools, and with the general expectation of the senior examining team. A significant shift in subject grade distribution compared to the

previous year will need to be explained. Perhaps there has been considerable growth in candidature, which has changed the overall level of achievement. Analysis of the relative performance of new schools may help in this case. A comparison of performance on those components with fixed grade boundaries from year to year, such as internal assessment, can also be a useful indicator. If the grade award meeting feels it necessary, it is possible to go back and review any grade boundary on any of the components for which boundaries have already been determined.

Once the grade award meeting participants, including appropriate IBO staff, are satisfied that the overall grade distribution is a fair reflection of candidate achievement in relation to the grade descriptors and results from previous years, individual school results are calculated and printed. A final check on the appropriateness of the results is then made by comparing awarded grades with predicted grades for a number of highly experienced schools in which it is known that teachers have a good record of familiarity with the required standards. Again, if there are significant discrepancies, further reviews can be conducted.

5.6.5 "At risking"

When the final results are generally deemed fair and correct, the senior examining team and other experienced examiners resolve outstanding issues relating to marking reliability. There may still be a few examiners whose work needs re-marking, or who have only recently been identified as suspect even though their moderation sample was satisfactory. The main area of re-marking, however, will concentrate on "at risk" candidates. Generally, these are candidates whose final grade is two or more grades worse than predicted and who are within two percentage marks of getting a better subject grade. Given that there is an error of measurement in marking and moderation, the accuracy of marking in such borderline cases needs to be confirmed, and all the externally marked components of these candidates will be re-marked. However, if it becomes clear that a school has been significantly over optimistic in many of its predicted grades for the subject and level, then less attention is likely to be paid to further re-marking for that school.

A second, much smaller, category of "at risk" candidate involves those candidates who are only one grade below prediction, and within two marks of achieving that predicted grade, where at least one externally marked component has been marked by an examiner whose marking was not fully satisfactory but who was not re-marked.

Ideally, all candidates within two marks of subject grade boundaries would be reviewed to check that they have received the correct grade. However, resource constraints do not permit this to take place and so attention is focused on those categories of candidate most likely to have suffered disadvantage from inaccuracy in marking and moderation.

5.6.6 Support personnel

Grade award meetings are supported by a large number of IBO staff, including SAMs/CAMs, assessment staff and examinations administration staff. There is also a large number of temporary assistants who receive, sort, check and move around all the candidate scripts received at IBCA. One of the most important duties of the temporary staff is to check every candidate script to make sure that the examiner has marked all the responses, awarded marks that are within the possible range for each question/task, and has correctly added up, transcribed and entered the total mark.

To help make the procedures of grade award meetings clearer to those not directly involved, teacher observers are invited to attend, in the expectation that they will report back to colleagues on their experience.

5.7 The final award committee

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

- cases of individual candidates affected by disability, accident or illness, which cannot be processed according to normal procedure
- cases of unforeseen circumstances affecting candidates
- cases of alleged malpractice
- policy recommendations from the special educational needs committee on assessment arrangements for candidates with special educational needs.

The final award committee is chaired by the chair of the examining board, and consists of a small number of other chief examiners and senior IBO staff. As for the grade award meetings, an observer from a school is invited to attend.

The committee follows established policy and precedent in dealing with candidates affected by adverse circumstances, with the intention of compensating where possible for any disadvantage suffered by the candidate. Sometimes this involves a slight increase in mark, and sometimes compensation for missing a whole examination component. In the latter cases, a statistical "missing mark procedure" can be employed to calculate a likely mark for the missing component based on marks gained in the other components compared to the average performance of candidates as a whole. There are certain conditions under which such compensation can be applied, and these conditions are part of the general regulations published in the *Vade Mecum*, the procedures manual for the formal DP assessment system.

Cases of maladministration by schools that have abused deadlines and/or procedures are considered by the committee. For serious cases that involve a major threat to the security and integrity of the examinations, or for repeated maladministration, it is possible that a school's authorization might be withdrawn.

Much of the committee's deliberations revolve around allegations of malpractice by candidates. Such allegations may be raised by the school itself, in relation to a candidate's conduct in an examination for example, or may be raised by examiners who feel they can detect plagiarism or collusion in candidates' work. Another possible form of malpractice is for candidates to inform other candidates in other parts of the world of the contents of question papers, prior to their examination. This is possible because of the time zone differences, which prevent examinations from being taken simultaneously on a global basis. Examiners are asked to be alert to detecting possible cases, and the IBO is moving towards greater use of regional variants of question papers in order to reduce the possibilities for such sharing of information between candidates.

In each case of alleged malpractice, as much background information as possible is sought and statements from the school and candidates involved are obtained before the committee makes a final judgment. Candidates found guilty of malpractice will be denied a result in that subject and consequently cannot be awarded the diploma.

5.8 Publication of results

Diploma and certificate results are published to schools and university admission systems on 5 January and 5 July each year for the two examination sessions. The results are sent electronically, as are many other administrative processes relating to the examination system, such as candidate registration and most mark entry. Candidates are issued with a numeric grade from 1 to 7 for each subject entered and candidates following the full Diploma Programme will

also receive letter grades for TOK and the extended essay, together with a total diploma points score. A bilingual diploma can be awarded to a candidate who:

1. takes two languages A1
2. takes a language A1 and a different language A2,

   or

3. takes examinations, or submits an extended essay, in at least one of the subjects from group 3 or group 4 in a language other than that taken as their language A1.

The maximum possible diploma points total, 45, is achieved by very few candidates, with about 5% of all full diploma candidates gaining more than 40 points. The diploma pass rate has remained fairly stable at approximately 80% in recent years. In mid-February/mid-August formal documentation of the results is sent to schools.

5.9 Feedback and enquiries upon results

In recent years, the IBO has placed great emphasis on improving and strengthening the feedback it provides to schools and teachers relating to the DP assessment system and to candidate achievement. This has the double benefit of strengthening the support provided by summative assessment to classroom teaching and also clarifying the workings of the assessment system for the schools, teachers and candidates who use it. After each examination session, the examination papers and their associated markschemes are made available for schools to purchase. Subject reports are written by the senior examining team, covering all general aspects of candidate performance on each component, outlining where candidates performed well, where they seemed less capable, and making recommendations for improving the preparation of candidates. The reports also contain the grade boundaries applied to each component. Subject reports are made available directly to teachers through the online curriculum centre (OCC), a web site devoted to providing professional support for IB teachers.

Soon after the official results are released, electronic information is automatically made available to schools giving the moderated mark and grade for each component of the assessment for each candidate. Component marks and grades do not form an officially published part of candidate results for the reasons outlined in section 3.2. These are provided to schools as useful feedback information, indicating the relative strengths and weaknesses of their candidates across the different components of a subject assessment model. Schools are also informed of how moderation affected the marks awarded by their teachers for each internal assessment component. In addition, an internal assessment feedback form prepared by the moderator, and giving brief information about the strengths of the internal assessment carried out for each subject and what could be improved, is sent electronically to schools.

As well as these more general forms of feedback, schools are able to use the enquiry upon results service to follow up more specific cases. There are three fee-based categories of enquiry. Category 1 is a re-mark of a candidate's externally marked work for a subject, if a school or candidate feels that the result is not a fair reflection of performance. If the subject grade is changed as a consequence of this re-mark, then no fee is charged. Subject grades may be raised or lowered as a result of a re-mark.

Category 2 is for schools to request the return of copies of the externally marked work of a complete school group of candidates for a given component, which may be an examination or a non-examination component. This allows teachers to see how a piece of work from their class of candidates was marked. Examiners are encouraged to write brief, constructive comments on the candidates' work when marking it. These help moderators and senior examiners who may review their marking later, but also provide further useful feedback to teachers. Although the intent of this category of enquiry is to provide feedback to teachers to inform their future teaching, it is also possible for a school to follow such an enquiry with a request for a re-mark through category 1.

Category 3 is a report written by the moderator explaining, in considerably more detail than is possible on the internal assessment feedback form mentioned above, why a teacher's internal assessment marks for each criterion were confirmed or adjusted. Further detailed information will also be supplied regarding the suitability of the actual tasks used for the internal assessment. It is intended that schools should use this service when they have already received the internal assessment feedback form and moderated mark adjustment data, but feel that further information is needed to understand why a moderation adjustment has been applied.

All the sources of feedback described in this section are made available to teachers and schools with the express aim of using the formal DP assessment system to make classroom teaching more effective and to enhance student learning.

Appendix A

Validity, reliability and generalizability—some further background

This section is included for those readers who require a more detailed description of the fundamental assessment concepts of validity and reliability.

A.1 Validity

The classic work on validity is that of Messick (1989). He recognized the need to treat validity as a unitary concept with construct validity as the underlying theme. Construct validity refers to the adequacy of an assessment instrument in reflecting an underlying domain or skill, so that no significant elements of the construct are under-represented, and no additional variables extraneous to the construct are included. Every task of an assessment instrument should be addressed solely and purely at the underlying construct. A psychometric view of a construct, such as writing, is likely to be considerably narrower than a performance assessment view of that same construct, which is likely to place greater emphasis on creativity and complex productive skills. In performance assessment, it is unlikely that every facet of a perceived construct can be represented in the assessment, and sampling of the domain content will be necessary. Construct validity is generally evaluated qualitatively by expert judgment.

In Messick's view, if construct validity is present, then the other types of validity are likely also to be embraced. Predictive validity refers to the ability of an assessment instrument to predict performance on future assessments of a similar nature. Concurrent validity refers to how well performance on an assessment instrument correlates with performance on another assessment instrument designed to measure the same construct. Concurrent validity and predictive validity clearly have much in common, and can be measured statistically by correlation. If there is a shared construct adequately represented in the different forms of assessment instrument, then it would seem logical that both concurrent and predictive validity will be high. However, problems can arise with predictive validity where significant selection occurs in the progression of the student population from one level of assessment to a future level, where different educational experiences take place between one assessment level and the next, and where there might be development or elaboration of the construct from one level of assessment to the next. Content validity, as its name suggests, relates to the actual knowledge and skill content of each assessment item/task, and whether this is appropriate for the stated construct. Content validity is generally measured by professional judgment, and obviously possesses considerable overlap with construct validity.

Messick also argued that the validity of assessments depends partly on the interpretation of assessment results and the intended and unintended social consequences of these. If assessment results are put to inappropriate use, then the validity of the assessment is reduced, also implying a failing in the construct validity of the assessment. Moss (1992, quoted in Gipps, 1994) stated that "the essential purpose of construct validity is to justify a particular interpretation of a test score by explaining the behaviour that the test score summarizes". This represents a considerable shift from the traditional psychometric practice of treating validity solely in terms of construct and content-related evidence, together with correlation studies of performance on tests intended to measure the same construct. The relatively recent notion of consequential validity has obvious implications for backwash on teaching, and has been a major driving force in the development of performance assessment in some parts of the world. A recognition that high-stakes assessment tends to have major structural influences on education systems (for example, Frederiksen and Collins, 1989; Broadfoot, 1996), together with the increasing recent impact of legislation on testing in the USA, have supported the importance now attached in validity studies to the social consequences of assessment.

A.2 Reliability

There are two main approaches to reliability measurement, approaches that are intrinsic and extrinsic to the assessment instrument. The intrinsic approach involves a consideration of the properties of the assessment instrument itself. Test-retest and parallel-forms reliability are both intrinsic methods, concerned with evaluating the stability of student responses. In a test-retest procedure, the same test is given to students, with only a short time interval between, and with the results unknown to the students, to see how similar the student responses are on the two occasions. For psychometric tests, highly similar responses cannot be taken for granted because many test items are designed to operate at students' maximum level of uncertainty (that is, to give the greatest discrimination between students). In performance assessment, this procedure is less applicable because student performance on the second instance of the assessment is likely to be highly dependent on the way they responded to the first instance, artificially inflating the measure. Parallel-forms reliability considers the degree of similarity in response to two different versions of the "same" test. For psychometric tests, this often involves the preparation of another test from a common bank of test items, given to the same students. Allowance would have to be made for slightly different item performance statistics of the particular samples of items that comprise each test.

Taking examinations as a typical example of performance assessment, the examination papers in the same subject set for different examination sessions could constitute parallel forms, although allowance would again have to be made for there being no expectation that identical levels of performance would be rewarded with identical marks on the two sessions. Matching would have to be done at a higher level of reporting, either grades or marks that have been processed in some way. Although test-retest and parallel-forms methods of calculating reliability are robust and provide good quality data, these methods are not commonly used because of practical demands and heavy resource implications (Feldt and Brennan, 1989). A much more accessible and widely used intrinsic measure of reliability depends on comparing student performance on one part of a test with performance on the rest of it, in other words, considering the internal consistency of a test. This is regarded as appropriate because all items (questions) in the test are designed to assess the same attribute or skill—each item is only included if student performance on that item correlates well with student performance on the whole test. This requires only a single administration of the test to a single group of students.

Dividing up the test can be done in a variety of ways, comparing the first half of the test with the second half, or odd-numbered items with even-numbered items (split-half method), or more sophisticated breakdowns such as statistically compounding all possible ways of dividing the test in two. In effect, the two halves of a single test are treated as parallel forms with the shortcoming that there is no separation of the assessment events over time.

Such measures of internal consistency have become commonly used as the principal reliability measures for standardized tests. Unfortunately, internal consistency is by no means an expected characteristic of most performance assessments, such as examinations. There is no expectation that student performance on one task should equate to their performance on another task in the same assessment instrument, although in practice there often is a fairly strong relationship. In fact the task may well have been deliberately designed to assess performance on quite different types of skill within the broad domain of the subject area being examined. As a result, there is little point in applying internal consistency measures of reliability to anything other than standardized tests. It is of more than passing interest to note that when different types of reliability measure (test-retest, parallel-forms and internal consistency) have been applied to the same test, significant discrepancies between the different measures are often found. For example, a study on the Iowa test of basic skills (ITBS, 1986, quoted in Wood, 1991) found that changes in student performance on different days gave a significantly reduced reliability measure compared to that given by internal consistency.

A.3 Marker reliability

The other aspect of reliability, the extrinsic one of marking consistency, is, however, very much an area of major concern for performance assessment. Standardized objective tests do not permit any variation in marks awarded to different responses, each one being either clearly correct or incorrect. However, performance assessments, which rely heavily on marker judgment, are vulnerable to differences of opinion on the merit of a given piece of student work. Such concerns divide into two: intra-marker reliability (how consistent is a given marker in allocating the marks to the same piece of work on different occasions?) and inter-marker reliability (how similar are the marks allocated to a given piece of work by two different markers?).

There has been little recent published research on the reliability of marking in examination systems dependent on marker judgment. However, earlier research found low measures for both inter- and intra-marker reliability (University of Cambridge Local Examinations Syndicate, 1976; Willmott and Nuttall, 1975; Nuttall and Willmott, 1972; Murphy, 1978 and 1982). The research shows that inter- and intra-marker reliability levels are affected by the nature of the individual assessment task. Short-response questions and highly structured responses to analytical questions can be marked more reliably than open-ended essay tasks or artistic performances/products. Marker training and the provision of detailed markschemes (rubrics) have been shown to raise marking reliability to high levels (Brown, 1992; Shavelson *et al*, 1992).

The issue of reliability is one that separates psychometric and performance assessment fundamentally. For psychometric tests, there is a readily available measure of reliability, or at least internal consistency, which carries a high value in the marketability of a test, and has long-term value because of the extended shelf life of a standardized test. For performance assessment instruments, internal consistency is not an intended outcome and the emphasis is on marking reliability. However, there is little point in going to great lengths in establishing a marker reliability measure for a given assessment instrument such as an examination, which has a one-off use and is then immediately discarded. Instead, emphasis is placed on quality control mechanisms in the system-wide processes of marking, in order to raise marker consistency to as high a level as possible for each different construction of a given assessment instrument.

Such quality control mechanisms are an important feature of all performance assessment systems, whether in the USA, Europe, Australasia or other parts of the world. In some systems all markers are collected together to mark in one place, under the watch of a senior marker who makes constant checks of their work. Sometimes, reliability can be improved further by requiring each marker to work on only one question or task out of the whole assessment instrument. Alternatively, markers may work individually but send a sample of their marking to a senior marker for re-marking. On the basis of a comparison of the set of paired marks, a statistical adjustment (or moderation adjustment) can be applied to the original marker's marking. A third approach is to have each student's work marked independently by two markers; if they agree within certain limits, the average mark of the two is awarded, but if there is a major disagreement the work is marked for a third time by a senior marker for arbitration. Fourth, in some systems markers are given deliberately random allocations of student work, on the understanding that the resulting distributions of marks for each marker should then be highly similar. Statistical adjustments are applied to each marker's distribution of marks to make that distribution match the one obtained by the most senior marker in the team.

These quality control procedures have different resource requirements and vary in feasibility according to context. They also vary in their relative advantages and disadvantages, but all are aimed at resolving the problem of ensuring satisfactory levels of inter- and intra-marker reliability.

A.4 Grading systems and reliability

A feature of many performance assessment systems at the upper-secondary school level is that they are made up of a number of different assessment instruments, or components, each designed to assess the sometimes quite different sets of knowledge and skills covered within a course of study. This is particularly true of high-stakes performance assessment. The use of a number of independently marked assessment instruments, taken by the students on different occasions, reduces the impact of marker variability and student variability, as well as providing the opportunity to build up a composite picture of student achievement in a variety of contexts. Additionally, such compound performance assessments are often reported on a much reduced scale of possible outcomes, for example, a two-state pass or fail, or pass, fail, merit and distinction, or a grading system consisting of a limited number of grades, for example A to G or 1 to 9.

The reason for this approach is that the reliability of a student result reported on a limited range of possible grades and based on the combination of a number of assessment instruments will be considerably higher than the reliability of the mark given on a single assessment instrument. There is an inevitable loss of information in reducing a student's reported result from one on a total scale of, for example, 400 marks to an outcome from a range of, say, five grades. However, when reporting such grade outcomes, the reliability can be just as high as that of the mark outcome from a standardized objective test. The much-reduced level of discrimination between students in reducing from a 400-point scale to a five-point scale is not a major concern, because fine discrimination between students is not regarded as a primary aim of performance assessment.

Additional devices to improve the reliability of final outcome in a graded performance system can include further checks and further re-marking by the most senior markers, of the work of candidates whose final compounded mark puts them close to the overall boundary between one grade and the next. In such cases, a small error of measurement for each student can have a major impact on the final result and the student's future life chances, in a high-stakes environment. The fewer the number of final grades available, the greater the impact of a student receiving the wrong grade. Also, in some assessment systems students can ask for a further re-mark of their work after the results have been issued, if they feel that the result is an unfair reflection of their achievement.

In the context of grading systems, there is another reliability issue that applies to some performance assessment systems, that of reliability of grading. There are some performance assessment instruments, for example, examination question papers, that will vary slightly in difficulty as different versions are prepared for each use, despite concerted efforts to make each version as comparable as possible in difficulty. Since the grades ultimately awarded at the conclusion of the assessment are intended to represent consistent standards of achievement, whereas the marks gained may be slightly harder or easier to achieve on different examples of the assessment instrument, it follows that the scale for converting marks into grades may need adjustment on each occasion. This is done by setting grade boundaries that define the minimum mark required to achieve each grade. In some systems, these are determined judgmentally by subject experts, who review the overall body of student work on each occasion. To enhance the reliability of grading, the process of establishing boundaries is arrived at by consensus among a group of experienced markers, who have previous examples of grade-boundary work to act as points of comparison, and statistical evidence to support their decision. However, because of the personal, subjective nature of each expert's contribution to the judgment, there will inevitably be some slight variability in the agreed judgments (Cresswell, 1996).

An alternative approach sometimes adopted is to adjust student grade distributions in any given subject to match the distribution of those same students in other subjects. This is a circular, iterative statistical process, known as inter-subject scaling, and can only be applied in the context of an overall fixed distribution of grades across all subjects. This approach is an

example of performance assessment feeding into a pre-determined grade distribution, something usually associated with standardized testing. In such a system, grades cannot be taken as representing set standards of achievement, but instead indicate each student's ranking within that particular cohort. Since there is no expressed intention of reporting the same grade for the same level of achievement on different occurrences of the assessment, the concept of reliability of grading is not applicable in such a system, though reliability of marking is still of course important.

A.5 Generalizability

The value of a score or grade awarded as the outcome of an assessment system lies in its generalizability—how much does the result tell us about what a student could do on other occasions, in different, but related, contexts? Any assessment can only be based on a sample of possible behaviour, with the intention of generalizing from achievement on this sample to achievement on the whole universe of a particular field of behaviour (Nuttall, 1987). To allow generalization, the whole field of behaviour (or construct) must be defined, and assessment must be reliable. Generalizability therefore depends on both validity and reliability, linking the two.

One of the biggest challenges to generalizability is the role of context. There is a parallel here between assessment and learning. A traditional view of learning might have suggested that complex concepts applicable to many areas should most economically be presented in de-contextualized, abstract form, and broken down into separate pieces. However, more recent understanding of the way learning develops (for example, Wood, 1998; Murphy, 1999; Shepard, 1992), emphasizes the complex nature of learning, the need for abstract procedures to be thoroughly connected to concrete problems and activities, and the essentially social nature of learning.

Such a situated approach to learning should inevitably impact on both the conduct of assessment, itself a social activity, and inferences made from it, if a supportive relationship is to be maintained between assessment and learning. For standardized objective tests, the range of context to which a test applies should be stated. For performance assessment, which is generally based on a broader view of underlying construct, increasing the number and spread of different kinds of task in the overall assessment is the most effective way of enhancing generalizability (Linn, 1993). Linn *et al* (1991) have also suggested that just as the concept of validity has been expanded to include consequences of the assessment (see appendix A.1), so there is an argument that the concept of reliability should also be expanded to include inferences made from test score onto a broader domain of achievement. If reliability were viewed in this way, then the internal consistency measures so often used for standardized multiple-choice tests, would be inadequate. In the context of standardized multiple-choice tests, reliability and validity are focused internally onto the test itself, whereas for performance assessment the two concepts have a more outward looking perspective, focusing on the utility of assessment and its impact on other aspects of education. Generally speaking, results from performance assessment are likely to be more generalizable (or less context specific) than results from psychometric testing.

Appendix B

IB Diploma Programme assessment policy

1. All assessment in Diploma Programme subjects should relate directly to the course of study and its objectives via a policy, as far as it is practicable, of discrete testing within each assessment environment (written papers/internal assessment and so on). A full range of assessment techniques should be used that reflect the international breadth of the IBO. The same assessment methodology should apply to related subjects but any substantial difference in the nature of higher level and standard level in a subject should be mirrored in their respective assessment models.
2. Diploma Programme assessment and grading procedures should ensure parity of treatment for all candidates irrespective of school, subject, response language or examination session. All grading and assessment judgments should be based on evidence and should not be subject to any form of bias.
3. All courses should normally have either three or four separate assessment components. Where appropriate, these components will include internal (school based) assessment as well as external assessment. No individual assessment component should normally be worth less than 20% or more than 50% of the overall assessment, and internally assessed components should in total contribute no more than 50% of the overall assessment. The balance between internal and external assessment must be such as to ensure that all the objectives of the course are adequately and appropriately assessed.
4. The duration of written examinations must not exceed five hours in total at higher level and three hours at standard level. No single written examination paper should be longer than three hours. Wherever possible, examination paper durations should be less than the prescribed maximum, as long as the examinations still provide for valid and reliable assessment. This restriction on duration is particularly relevant in those subjects where internal assessment or other externally marked components form a significant part of the overall assessment model.
5. The marking of teachers and examiners will be moderated using a mark/re-mark model followed by a statistical comparison to generate a moderation equation. There will be no cross-component moderation. All such re-marking will be based on identical assessment criteria to the original marking and will be based on sample work sent to an examiner acting as moderator.
6. Internal assessment should primarily address those skills and areas of understanding that are less appropriately addressed through external examination papers; it should not be treated as another means for candidates to demonstrate, in a different context, what they could also do in an examination. There should be no undue duplication of skills assessed in both internal assessment and external examination.
7. Internal assessment should not be used as a tool for monitoring syllabus coverage, but should be focused on assessing student learning of particular skills. Where necessary, breadth of syllabus coverage should be assessed within external examinations.
8. Internal assessment tasks should not duplicate the kind of work that is carried out for extended essays in the same subject.
9. Wherever possible, internal assessment tasks should become an integral part of normal classroom teaching (and/or homework) for that subject. They should not be "add-on" activities. The work carried out for internal assessment is meant to be part of each student's learning experience.
10. For internal assessment marks to make a reliable contribution to a candidate's subject grade, the work that contributes at least half of the total internal assessment mark must be susceptible to moderation. This is a minimum, it being preferable wherever possible for all of the work that gives rise to the internal assessment mark to be available for moderation.

11. Where different internally assessed tasks are carried out over a prolonged period within a Diploma Programme course (to make up a portfolio of work, for example) allowance must be made for student improvement over this period. Thus the final internal assessment mark should reflect a student's best level of performance during the course and not be merely an average of performance over the whole course.
12. Although the internal assessment may contribute from 20% to 50% towards any single subject result, the higher values in this range should only be used where there are particular grounds for giving a high weighting to internally assessed work.
13. Internally assessed work must be produced under conditions that are well documented and common to all schools for each course. In particular, the role of collaborative work, the degree of assistance that teachers can provide, the extent to which students can use external resources, and the permitted amount of redrafting of work, must be fully described.
14. The quantity of internally assessed work specified for a course must be no more than the minimum needed to satisfy its aims. Defined word limits should be given where possible for internally assessed tasks. The maximum word limit should be no more than is necessary to complete the task.

References

Assessment Reform Group (1999) *Assessment for Learning: Beyond the Black Box*, Cambridge: University of Cambridge School of Education.

Biggs, J (1998) "Assessment and classroom learning: the role of summative assessment", *Assessment in Education*, 5(1): 103–110.

Binet, A and Simon, T (1905) "Methodes nouvelles pour le diagnostique du niveau intellectuel des anormaux", *L'annee psychologique*, II: 245–336.

Black, P (1998) *Testing: Friend or Foe? Theory and Practice of Assessment and Testing*, London: Falmer Press.

Black, P (1999) "Assessment, Learning Theories and Testing Systems", in Murphy, P (1999) *Learners, Learning & Assessment*, London: Paul Chapman Publishing in association with The Open University.

Black, P and Wiliam, D (1998a) "Assessment and Classroom Learning", *Assessment in Education,* 5(1): 7–71.

Black, P and Wiliam, D (1998b) *Inside the Black Box: Raising standards through classroom assessment*, London: School of Education, King's College.

Black, PJ (1993a) "Assessment policy and public confidence: Comments on the BERA Policy Task Group's article, *Assessment and the improvement of education*", *The Curriculum Journal*, 4(3): 421–7.

Black, PJ (1993b) "Formative and summative assessment by teachers", *Studies in Science Education*, 21: 49–97.

Bloom, BS, Englehart, MD, Furst, EJ, Hill, WH and Krathwohl, DR (1956) *Taxonomy of Educational Objectives. Handbook 1: Cognitive Domain*, New York: David McKay.

Broadfoot, P (1996) *Education, Assessment and Society: A Sociological Analysis*, Buckingham: Open University Press.

Brown, M (1988) "Issues in formulating and organising attainment targets in relation to their assessment", in Torrance, H (ed) *National Assessment and Testing: A Research Response*, London: British Educational Research Association.

Brown, M (1992) "Elaborate nonsense? The muddled tale of SATs in mathematics at KS 3", in Gipps, C (ed) *Developing Assessment for the National Curriculum*, London: ULIE/Kogan Page.

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: Kogan Page.

Cannell, JJ (1988) "Nationally normed elementary achievement testing in America's public schools: How all 50 states are above the national average", *Educational Measurement: Issues and Practice*, 7(2): 5–9.

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 and New York: Wiley, 57–84.

Feldt, LS and Brennan RL (1989) "Reliability" in Linn, RL (ed) *Educational Measurement* (3rd edition), Washington, DC: American Council on Education/Macmillan Series on Higher Education.

Frederiksen, J and Collins, A (1989) "A systems approach to educational testing", *Educational Researcher*, 18(9): 27–32.

Frith and Macintosh (1984) *A Teachers' Guide to Assessment*, Cheltenham: Stanley Thomas.

Gardner, H (1983) *Frames of mind*, New York: Basic Books.

Gardner, H (1999) "Assessment in Context", in Murphy, P *Learners, Learning & Assessment*, London: Paul Chapman Publishing in association with The Open University.

Gipps, C and Murphy, P (1994) *A Fair Test? Assessment, Achievement and Equity*, Buckingham: Open University Press.

Gipps, CV (1994) *Beyond Testing: Towards a Theory of Educational Assessment*, London: Falmer Press.

Gipps, CV and Stobart, G (1993) *Assessment: A Teachers' guide to the Issues* (2nd edition), London: Hodder and Stoughton.

Glaser, R (1963) "Instructional technology and the measurement of learning outcomes: Some questions", *American Psychologist*, 18: 519–21.

Goldstein, H (1996a) "Group differences and bias in assessment", in Goldstein, H and Lewis, T (eds) *Assessment: Problems, Developments and Statistical Issues,* Chichester and New York: Wiley, 85–93.

Goldstein, H (1996b) "The statistical analysis of institution based data", in Goldstein, H and Lewis, T (eds) *Assessment: Problems, Developments and Statistical Issues*, Chichester and New York: Wiley, 135–44.

Harlen, W (ed) (1994) *Enhancing Quality in Assessment*, BERA Policy Task Group on Assessment, Paul Chapman Publishers.

Hieronymus, AN and Hoover, HD (1986) *Iowa Tests of Basic Skills: Manual for School Administrators, Levels 5–14*, Chicago: Riverside Publishing Company, 156–8.

Hill, I (2002) "The history of international education: an International Baccalaureate perspective", in Hayden, M, Thompson, J and Walker, G (eds) *International education in practice: dimensions for national and international schools*, London: Kogan Page.

Hoffmann, B (1962) *The Tyranny of Testing*, New York: Crowell–Collier.

Hughes, DC, Keeling, B and Tuck, BF (1983) "Are untidy essays marked down by graders with neat handwriting?", *New Zealand Journal of Educational Studies*, 18: 184–6.

Humphreys, LG (1986) "An analysis and evaluation of test and item bias in the prediction context", *Journal of Applied Psychology*, 71: 327–33.

International Baccalaureate Organization (1999) Diploma Programme *Language A1 guide*, Cardiff: IBO.

International Baccalaureate Organization (2001a) Diploma Programme *History guide*, Cardiff: IBO.

International Baccalaureate Organization (2001b) Diploma Programme *Biology guide*, Cardiff: IBO.

International Baccalaureate Organization (2003a), *Perceptions of the International Baccalaureate Diploma Programme*, Cardiff: IBO.

International Baccalaureate Organization (2003b), *Academic Honesty: guidance for schools*, Cardiff: IBO.

Iowa Tests of Basic Skills (1986) *Manual for School Administrators: Levels 5–14*, Chicago: Riverside Press.

Kingdon, M and Stobart, G (1987) *The Draft Grade Criteria: A Review of LEAG Research*, LEAG Discussion Paper.

Lambert, D and Lines, D (2000) *Understanding Assessment*, London: RoutledgeFalmer.

Linn, MC (1992) "Gender differences in educational achievement", in *Sex Equity in Educational Opportunity, Achievement, and Testing*, Princeton, NJ: Educational Testing Service.

Linn, RL (1993) "Educational assessment: Expanded expectations and challenges", *Educational Evaluation and Policy Analysis*, 15: 1.

Linn, RL, Baker, E and Dunbar, S (1991) "Complex, performance-based assessment: Expectations and validation criteria", *Educational Researcher*, 20(8): 15–21.

Messick, S (1989) "Validity", in Linn, R (ed) *Educational Measurement* (3rd edition), American Council on Education, Washington: Macmillan.

Moss, PA (1992) "Shifting conceptions of validity in educational measurement: Implications for performance assessment", *Review of Educational Research*, 62(3), 229–58.

Murphy, P (1999) *Learners, Learning & Assessment*, London: Paul Chapman Publishing in association with The Open University.

Murphy, RJL (1978) "Reliability of marking in eight GCE examinations", *British Journal of Educational Psychology*, 48: 196–200.

Murphy, RJL (1982) "A further report of investigations into the reliability of marking of GCE examinations", *British Journal of Educational Psychology*, 52: 58–63.

Nuttall, D (1987) "The validity of assessments", *European Journal of Psychology of Education*, 11(2): 109–18.

Nuttall, DL and Willmott, AS (1972) *British Examinations: Techniques of Analysis*, Slough: NFER Publishing Co.

Orr, L and Nuttall, D (1983) *Determining standards in the proposed system of examining at 16+*, Comparability in Examinations Occasional Paper 2, London: Schools Council.

Peterson, ADC (1971) *New Techniques for the assessment of Pupils' Work*, Strasbourg: Council of Europe.

Peterson, ADC (2003) *Schools Across Frontiers: The Story of the International Baccalaureate and the United World Colleges*, (2nd edition), Chicago: Open Court Publishing Company.

Popham, J (1987) "The merits of measurement-driven instruction", *Phi Delta Kappa*, May, 679–82.

Popham, WJ (1978) *Criterion-referenced Measurement*, Englewood Cliffs, NJ: Prentice Hall.

Resnick, L (1989) "Introduction" in Resnick, L (ed) *Knowing, Learning and Instruction*. *Essays in honour of R Glaser*, New Jersey: Lawrence Erlbaum Associates.

Resnick, LB and Resnick, DP (1992) "Assessing the thinking curriculum: New tools for educational reform", in Gifford, B and O'Connor, M (eds) *Changing Assessments: Alternative Views of Aptitude, Achievement and Instruction*, London: Kluwer Academic Publishers.

Sadler, R (1987) "Specifying and promulgating achievement standards", *Oxford Review of Education*, 13: 2.

Sadler, R (1998) "Formative assessment: revisiting the territory", *Assessment in Education*, 5(1): 77–84.

Satterley, D (1994) "The quality of external assessment" in Harlen, W (ed) *Enhancing Quality in Assessment*, Paul Chapman Publishers.

SEC (1984) "The Development of Grade-Related Criteria for the GCSE: A briefing Paper for Working Parties", London: SEC.

Shavelson, R, Baxter, G and Pine, J (1992) "Performance assessments: Political rhetoric and measurement reality", *Educational Researcher*, 21: 4.

Shepard, L (1991) "Psychometricians' beliefs about learning", *Educational Researcher*, 20: 7.

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 and Dordrecht: Kluwer, 301–28.

Smith, D and Tomlinson, S (1989) *The School Effect*, London: Policy Studies Institute.

University of Cambridge Local Examinations syndicate (1976) *School Examinations and their Function*, Cambridge: UCLES.

Vygotsky, LS (1962) *Thought and language*, New York: Wiley.

Vygotsky, LS (1978) *Mind and society: The development of higher psychological processes*, Cambridge: Harvard University Press.

Wiggins, G (1989) "Teaching to the (authentic) test", *Educational Leadership*, 46(7): 41–7.

Wiliam, D and Black, PJ (1996) "Meanings and consequences: a basis for distinguishing formative and summative functions of assessment", *British Educational Research Journal*, 22(5): 537–48.

Willmott, AS and Nuttall, DL (1975) *The Reliability of Examinations at 16+*, London: Macmillan.

Wood, D (1998) *How Children Think and Learn: The Social Contexts of Cognitive Development* (2nd edition), Oxford: Blackwell.

Wood, D, Bruner, JS and Ross, G (1976) "The role of tutoring in problem solving", *Journal of Child Psychology and Psychiatry*, 17: 89–100.

Wood, R (1991) *Assessment and Testing: A survey of research*, Cambridge: Cambridge University Press.

