

# Principios y práctica del sistema de evaluación del Programa del Diploma





Programa del Diploma

# Principios y práctica del sistema de evaluación del Programa del Diploma

**Organización del Bachillerato Internacional**

Buenos Aires

Cardiff

Ginebra

Nueva York

Singapur

## Principios y práctica del sistema de evaluación del Programa del Diploma

Versión en español del documento publicado en septiembre de 2004  
con el título *Diploma Programme assessment: principles and practice*

Publicada en enero de 2005

por la Organización del Bachillerato Internacional  
Peterson House, Malthouse Avenue, Cardiff Gate  
Cardiff, Wales GB CF23 8GL  
Reino Unido  
Tel.: + 44 29 2054 7777  
Fax: + 44 29 2054 7778  
Sitio web: [www.ibo.org](http://www.ibo.org)

© Organización del Bachillerato Internacional, 2005

La Organización del Bachillerato Internacional es una fundación educativa internacional sin fines de lucro. Fue creada en 1968 y tiene sede legal en Suiza.

IBO agradece la autorización para reproducir en esta publicación material protegido por derechos de autor. Cuando procede, se han citado las fuentes originales y, de serle notificado, IBO enmendará cualquier error u omisión con la mayor brevedad posible.

El uso del género masculino en esta publicación no tiene un propósito discriminatorio y se justifica únicamente como medio para hacer el texto más fluido. Se pretende que el español utilizado sea comprensible para todos los hablantes de esta lengua y no refleje una variante particular o regional de la misma.

Los artículos promocionales y las publicaciones de IBO en sus lenguas oficiales y de trabajo pueden adquirirse a través del catálogo en línea, disponible en [www.ibo.org](http://www.ibo.org) al seleccionar **Publicaciones** en el menú de atajos. Las consultas sobre pedidos deben dirigirse al departamento de ventas en Cardiff.

Tel.: +44 29 2054 7746  
Fax: +44 29 2054 7779  
Correo-e: [sales@ibo.org](mailto:sales@ibo.org)

## Declaración de principios de IBO

La Organización del Bachillerato Internacional tiene como meta formar jóvenes solidarios, informados y ávidos de conocimiento, capaces de contribuir a crear un mundo mejor y más pacífico, en el marco del entendimiento mutuo y el respeto intercultural.

En pos de este objetivo, la Organización del Bachillerato Internacional colabora con establecimientos escolares, gobiernos y organizaciones internacionales para crear y desarrollar programas de educación internacional exigentes y métodos de evaluación rigurosos.

Estos programas alientan a estudiantes del mundo entero a adoptar una actitud activa de aprendizaje durante toda su vida, a ser compasivos y a entender que otras personas, con sus diferencias, también pueden estar en lo cierto.



# Índice

<b>1</b>	<b>Introducción y panorama general</b>	<b>1</b>
<b>2</b>	<b>Principios de la evaluación</b>	<b>3</b>
2.1	¿Por qué “evaluación”?	3
2.2	Formativa y sumativa: ¿cuáles son los fines de la evaluación?	4
2.3	Pruebas psicométricas y evaluación del rendimiento	5
2.4	Evaluación y aprendizaje	7
2.5	Evaluación normativa y evaluación por criterios	7
2.6	Validez y fiabilidad: panorama general	9
2.7	Sesgo	10
<b>3</b>	<b>La evaluación en el Programa del Diploma: objetivos y enfoques</b>	<b>13</b>
3.1	Apoyo a los objetivos curriculares	13
3.2	Fiabilidad de los resultados	14
3.3	Dimensiones internacional e intercultural	15
3.4	Capacidades cognitivas superiores	18
3.5	Gama de tareas de evaluación y de instrumentos de evaluación (componentes)	19
3.6	Papel de la opinión profesional	21
<b>4</b>	<b>Estructuras de evaluación del Programa del Diploma</b>	<b>23</b>
4.1	El currículo del Programa del Diploma	23
4.2	Modelos de evaluación y papel de la evaluación interna	26
4.3	Personal	28
<b>5</b>	<b>Procedimientos de evaluación del Programa del Diploma</b>	<b>30</b>
5.1	Preparación de exámenes	30
5.2	Los exámenes	32
5.3	Evaluación interna y otros componentes no de examen	33
5.4	El trabajo de corrección	34
5.5	Moderación	39
5.6	Totalización y concesión de calificaciones finales	45
5.7	Comité de la evaluación final	50
5.8	Publicación de resultados	51
5.9	Comentarios y consultas sobre los resultados	51
	<b>Apéndice A</b>	<b>53</b>
	Validez, fiabilidad y generalizabilidad: información adicional	53
	<b>Apéndice B</b>	<b>58</b>
	Política de evaluación del Programa del Diploma	58
	<b>Referencias</b>	<b>60</b>



## Introducción y panorama general

El objetivo de esta publicación es explicar los mecanismos y fundamentos del sistema de evaluación del Programa del Diploma del Bachillerato Internacional. El Programa del Diploma se imparte en más de 100 países, en colegios que representan una amplia variedad de contextos y tradiciones educativas. A algunos de estos colegios les resultarán familiares la filosofía y los enfoques que adopta la Organización del Bachillerato Internacional (IBO) para evaluar a sus alumnos, mientras que a otros puede que el sistema les parezca misterioso y confuso. Es importante que los profesores que preparan a alumnos para los exámenes del Programa del Diploma comprendan la naturaleza del sistema de evaluación al que van a someterlos, así como la filosofía educativa del programa en su conjunto. Casi todos los profesores del Programa del Diploma participan necesariamente en el proceso de evaluación, contribuyendo directamente a la evaluación final de sus propios alumnos. Además, muchos profesores del Programa del Diploma también toman parte como examinadores en la corrección del trabajo de otros alumnos, en la comprobación del trabajo de corrección de otros examinadores, o incluso en la redacción de exámenes (siempre que no vayan destinados a sus propios alumnos). Por todos estos motivos, comprender cómo encaja su propia aportación en el esquema general de la evaluación no puede sino resultar de utilidad para los profesores del Programa del Diploma.

Además de servir de documento de apoyo para profesores y coordinadores del Programa del Diploma, esta publicación también puede ser útil para los encargados de las admisiones a las universidades, para los directores de colegios, los órganos de gobierno de colegios, los examinadores y para los mismos alumnos. Aunque a muchas universidades de todo el mundo les resulta ya muy familiar recibir solicitudes de ingreso de alumnos que han cursado el Programa del Diploma y conocen bien sus cualidades (véase, por ejemplo, IBO, 2003a), quedan aún muchas más que desconocen la titulación del Diploma del BI. También puede que muchos alumnos y sus padres/tutores estén legítimamente muy interesados en comprender la naturaleza del sistema de evaluación que constituye una parte tan esencial de la participación en el Programa del Diploma del BI.

Esta publicación se centra en los aspectos formales de la evaluación en el Programa del Diploma, es decir, aquellos que contribuyen a la obtención del título. No tiene como finalidad discutir detalladamente el área igualmente importante de la evaluación formativa, por medio de la cual los profesores de clase pueden influir de manera más directa e inmediata en la evolución del aprendizaje de sus alumnos. Se aconseja a los profesores que deseen obtener más información sobre la evaluación formativa, que consulten obras como, por ejemplo, Black y Wiliam (1998a), Black y Wiliam (1998b), Assessment Reform Group (Grupo de reforma de la evaluación) (1999), y Sadler (1998), aunque la mayoría de libros estándar sobre evaluación educativa también tratan el tema.

A fin de comprender la naturaleza del sistema de evaluación del Programa del Diploma, es necesario disponer de información sobre el desarrollo histórico y teórico de la práctica de la evaluación. En esta publicación se ha abordado la cuestión examinando y describiendo brevemente dicho desarrollo. Muchas de las cuestiones importantes se han presentado de manera simplificada, sin embargo, las cuestiones históricas y conceptuales deben tratarse con más detalle debido a la importancia de su impacto en la práctica actual. Aquellos lectores que deseen obtener más información sobre las cuestiones y desafíos a los que se enfrentan los responsables de elaborar y aplicar los sistemas de evaluación, encontrarán en las obras que se citan en el texto un punto de partida adecuado para seguir investigando. Hay más detalles sobre los mecanismos y procedimientos de la evaluación en el Programa del Diploma en los muchos documentos de procedimiento interno elaborados por IBO pero, probablemente, sólo necesiten tantos detalles quienes se encarguen directamente del funcionamiento de la evaluación.

La siguiente sección de esta publicación, *Principios de la evaluación*, comenta los conceptos teóricos que han guiado el desarrollo de la evaluación educativa a gran escala a lo largo de las últimas décadas. Se introduce y explica terminología de uso común relativa a la evaluación. Se describen los diferentes antecedentes de los dos enfoques principales de la evaluación, las pruebas psicométricas y la evaluación del rendimiento, junto con los distintos fines para los que se utilizan los resultados de la evaluación. La sección concluye haciendo hincapié en los compromisos entre necesidades opuestas a los que ha de llegarse cuando se diseña un sistema de evaluación, dados los serios conflictos que pueden surgir entre fiabilidad, distintos aspectos de validez, equidad y viabilidad.

La sección 3, *La evaluación en el Programa del Diploma: objetivos y enfoques*, muestra cómo se tienen en cuenta los conceptos y principios descritos en la sección anterior a la hora de desarrollar los enfoques de la evaluación que se utilizan en el Programa del Diploma. Aunque la fiabilidad de los resultados finales y la validez de las evaluaciones son importantes en lo que se refiere al modo en que abordan las capacidades cognitivas a través de una amplia gama de tareas de evaluación distintas, ninguna de estas cuestiones resta valor al objetivo principal de idear un sistema de evaluación que apoye y fomente la enseñanza y el aprendizaje adecuados en el aula. Los otros principios importantes de la evaluación en el Programa del Diploma son: que los resultados deben basarse en la opinión profesional de examinadores supervisores, y que las evaluaciones deben reflejar las dimensiones internacional e intercultural del programa.

En la sección 4, *Estructuras de evaluación del Programa del Diploma*, se explican brevemente en términos generales las principales estructuras de evaluación, que son las siguientes:

- el modelo curricular, que describe los diferentes cursos objeto de evaluación, que se complementan para formar un programa de estudios equilibrado e integrado para cada alumno del Programa del Diploma
- los diferentes modelos de evaluación que se aplican a cada curso, en los cuales juega un papel importante la evaluación interna que lleva a cabo el profesor de clase
- los diferentes grupos de personas (personal de IBO, profesores, examinadores y demás personal académico) que desempeñan distintos papeles a la hora de proyectar las evaluaciones y de encargarse de su funcionamiento.

La sección final explica cronológicamente los *Procedimientos de evaluación del Programa del Diploma*, desde la preparación de las preguntas de los exámenes de cada convocatoria, pasando por el proceso de corrección y la concesión de calificaciones finales, hasta la comunicación de los resultados a los alumnos y el seguimiento que tiene lugar posteriormente. El período total que cubren todas estas actividades es de aproximadamente dos años, y puesto que hay dos convocatorias de exámenes cada año, en mayo y en noviembre, se produce un solapamiento considerable de las actividades correspondientes a cada convocatoria.

Esta publicación tiene como objetivo servir de documento de referencia a los diferentes tipos de lector mencionados al principio. Todos son parte interesada en la evaluación en el Programa del Diploma y quizás deseen consultar las distintas secciones de la publicación según vayan surgiendo problemas o preguntas. Se espera que, además de explicar cómo funciona la evaluación en el Programa del Diploma, esta publicación permita también comprender por qué la evaluación se lleva a cabo así. Los lectores principalmente interesados en el funcionamiento de la evaluación encontrarán de mayor interés las secciones 4 y 5, mientras que quienes estén interesados en los principios teóricos subyacentes encontrarán información en las secciones 2 y 3.

Los apéndices contienen más explicaciones sobre los diferentes aspectos de validez y fiabilidad, así como sobre la política de evaluación del Programa del Diploma que rige el desarrollo de los modelos de evaluación correspondientes a cada asignatura. Se invita a aquellos lectores que tengan preguntas que trasciendan el ámbito de la presente publicación a que se pongan en contacto con el personal de evaluación de IBO, dirigiéndose por correo electrónico a: [assessment@ibo.org](mailto:assessment@ibo.org).

## 2 Principios de la evaluación

### 2.1 ¿Por qué “evaluación”?

Para muchas personas, las palabras “evaluación”, “examen” y “test” tienen un significado similar y se utilizan más o menos indistintamente. A efectos de esta publicación, es necesario darles un significado más específico, que será siempre el siguiente:

**Test:** conjunto de muchas preguntas de respuesta corta (de elegir la respuesta/de opción múltiple o que debe responderse con sólo unas pocas palabras) que los alumnos deben contestar estando aislados y supervisados, en un tiempo determinado. A menudo se califica automáticamente.

**Examen:** conjunto de una o más tareas de distintos tipos (preguntas de respuesta corta, de respuesta extensa, de resolución de problemas, o analíticas; a veces tareas prácticas u orales) que los alumnos deben realizar estando aislados y supervisados, en un tiempo determinado. Generalmente calificado por un examinador.

**Evaluación:** término utilizado para referirse a todos los distintos métodos por los que pueden valorarse los logros del alumno. Entre los instrumentos de evaluación pueden incluirse tests, exámenes, trabajos prácticos extensos, proyectos, carpetas/portafolios, y presentaciones orales, algunos realizados durante un largo período de tiempo y a veces calificados por el profesor del alumno.

Con frecuencia se establece una distinción entre evaluación *sumativa*, cuyo objetivo es determinar el nivel de logro del alumno, generalmente al final del curso, y evaluación *formativa*, cuyo objetivo es identificar las necesidades de aprendizaje del alumno y que forma parte del proceso de aprendizaje en sí mismo. Aunque estas dos funciones parecen estar bastante diferenciadas, a menudo pueden utilizarse los mismos instrumentos de evaluación para ambos fines, estableciéndose la diferencia en cuanto al modo en que se interpretan y aplican los resultados de la evaluación (Black, 1993a; Wiliam y Black, 1996). Biggs (1998) también aclaró que no es conveniente considerar que la evaluación formativa y la sumativa sean mutuamente excluyentes. Ambas deben interactuar y servirse de apoyo una a la otra. En el contexto del Programa del Diploma, se prefiere el término *evaluación formal* para describir todos aquellos instrumentos de evaluación que contribuyen a la obtención del título. Algunos de ellos pueden utilizarse formativamente durante el curso, además de sumativamente hacia el final del mismo, enfoque que ha sido propuesto por otros autores (por ejemplo, Lambert y Lines, 2000, cap. 10).

La evaluación formal en el Programa del Diploma incluye algunos tests de opción múltiple en unas pocas asignaturas, y pruebas escritas en la mayoría de ellas, que los alumnos deberán hacer al final del curso de dos años de duración. También incluye otras tareas (redacciones, ensayos sobre investigación, trabajos escritos, entrevistas orales, investigaciones científicas y matemáticas, proyectos de trabajo de campo e interpretaciones artísticas) repartidas en distintas asignaturas, que los alumnos deberán realizar en diferentes momentos y condiciones a lo largo del curso. En la sección 3.5 se explicará el motivo de utilizar tantos instrumentos de evaluación distintos.

## 2.2 Formativa y sumativa: ¿cuáles son los fines de la evaluación?

La evaluación puede utilizarse con distintos fines. El fin que persiga un sistema de evaluación determinado influirá decisivamente en su estilo y formato. El fin principal de la evaluación formativa es proporcionar a profesores y alumnos información detallada sobre la naturaleza de los puntos fuertes y débiles de los alumnos, y contribuir al desarrollo de sus capacidades. Son especialmente útiles aquí los métodos de evaluación que implican interacción directa entre profesor y alumno. Se considera al profesor más como alguien que apoya el aprendizaje que como alguien que lo dirige (Vygotsky, 1962; Vygotsky, 1978), debiendo utilizar las tareas e instrumentos de evaluación para ayudar a trabajar al alumno en lo que Vygotsky denomina “zona de desarrollo próximo”. Se trata del tramo de logro que abarca desde lo que el alumno puede hacer por sí solo hasta lo que puede hacer con el apoyo del profesor. Es un concepto similar al de “andamiaje” creado por Wood et ál. (1976), según el cual el profesor proporciona el andamio para la construcción del aprendizaje, pero sólo el alumno puede llevar a cabo dicha construcción. El objetivo del profesor debe ser establecer evaluaciones formativas que planteen un reto del nivel exactamente adecuado para el alumno, y continuar ajustando dicho nivel según el alumno va progresando.

En la evaluación formativa, es más importante identificar correctamente los conocimientos, destrezas y grado de comprensión que deben desarrollar los alumnos, que medir exactamente el nivel de logro de cada alumno. La *fiabilidad* es por tanto una cuestión a tener mucho menos en cuenta en la evaluación formativa que la *validez* (véanse en la sección 2.6 las explicaciones de los términos “fiabilidad” y “validez”).

La evaluación sumativa se utiliza con fines bastante distintos, incluidos proporcionar información sobre los logros del alumno, permitir la certificación y selección de alumnos, así como servir de mecanismo de atribución de responsabilidades para hacer una valoración de profesores y colegios, y de fuerza para impulsar reformas del currículo. La utilización de la evaluación sumativa como mecanismo de atribución de responsabilidades, con objetivos como los de elevar el nivel y proporcionar información para determinar qué colegios y profesores son “los buenos”, es un tema controvertido (Gipps y Stobart, 1993; Goldstein, 1996b). La controversia se plantea en torno, por una parte, a la dificultad de establecer comparaciones justas entre profesores y colegios que puedan tener alumnos de muy distintas procedencias y estar enseñando en contextos muy diferentes y, por otra parte, a la dificultad de interpretar a qué pueda deberse el que aparentemente se eleven los niveles de rendimiento escolar. Esto puede reflejar que se han producido auténticas mejoras en la enseñanza y el aprendizaje, o que se han realizado mayores esfuerzos en la enseñanza centrada en los exámenes, llevando a un mayor descuido de otros aspectos de la educación.

La evaluación en el Programa del Diploma no tiene el papel formal de mecanismo de atribución de responsabilidades para juzgar el rendimiento del colegio, aunque los propios colegios y los padres de los alumnos puedan utilizar de este modo los resultados de la evaluación. IBO considera que corresponde en gran medida a cada colegio hacer una valoración de su propia eficacia. La manera en que los colegios autorizados implementan el programa se evalúa cada cinco años en un proceso denominado “evaluación del programa”. El proceso se basa primordialmente en el estudio que cada colegio hace de sí mismo. Rara vez se toma en cuenta el rendimiento escolar de los alumnos en el Programa del Diploma como factor decisivo para determinar la condición de colegio autorizado de IBO, aunque si el colegio incurre reiteradamente en la incorrecta administración de la evaluación o en conducta fraudulenta, pueda reconsiderarse la autorización concedida a dicho colegio para impartir el programa (IBO, 2003b). Las políticas que aplican los colegios para decidir qué alumnos pueden matricularse en el Programa del Diploma, y los contextos sociales y educativos en los que operan son tan diversos, que sería inadecuado juzgar la eficacia de un colegio únicamente con base en los resultados de los exámenes del Programa del Diploma.

Aunque en general se considera que el papel principal de la evaluación en el Programa del Diploma es certificar los logros de los alumnos, lo que desemboca en la mayoría de los casos en un proceso de selección para acceder a la universidad, también son importantes las demás aplicaciones de la evaluación sumativa. En la medida en que sea viable, la evaluación en el Programa del Diploma también puede servir

de herramienta importante para reforzar la enseñanza de los objetivos curriculares del programa: en realidad, dicha evaluación sólo puede ser válida si refleja adecuadamente esos objetivos. Un tercer fin, el de proporcionar información diferenciada sobre los logros del alumno (y, por tanto, sobre la eficacia del profesor) que contribuya al desarrollo profesional de los profesores, también contribuye de forma significativa a la influencia que el Programa del Diploma tiene en la educación de los alumnos. Para poder desempeñar con éxito estos diferentes papeles, y atender a las distintas exigencias que imponen en un sistema de evaluación, debe llegarse inevitablemente a ciertos compromisos en el diseño de la evaluación en el Programa del Diploma, tal y como se comenta en la sección 4.

Vale la pena señalar desde un principio, que cualquier análisis de los distintos sistemas de evaluación nacionales pondrá rápidamente de manifiesto la existencia de una gran variedad de técnicas y enfoques de evaluación. Todos estos sistemas tienen sus puntos fuertes y débiles en cuanto a consideraciones técnicas, de recursos y de tiempo, y en cuanto a su influencia en el sistema educativo al que pertenecen. Incluso si fuese posible, en un contexto dado, empezar completamente desde cero a diseñar un sistema de evaluación, no existe ninguna técnica que pueda aplicarse por ser considerada la mejor universalmente. En lugar de ello, las opciones que se eligen al diseñar sistemas de evaluación reflejan inevitablemente los valores o prioridades del contexto social más amplio en el que se lleva a cabo dicha selección (Cresswell, 1996; Broadfoot, 1996).

## 2.3 Pruebas psicométricas y evaluación del rendimiento

Existen principalmente dos tradiciones muy distintas en lo que atañe a la evaluación, que se utilizan ampliamente en los sistemas educativos nacionales en la actualidad. La primera de ellas, la psicometría, tiene sus raíces en el desarrollo de los tests de inteligencia, en París, a principios del siglo XX (Binet y Simon, 1905). Lo que se pretende con los tests psicométricos es utilizar cierto número de preguntas breves, cuidadosamente calibradas (generalmente preguntas de opción múltiple), para medir con precisión la aptitud o el potencial del alumno en un área concreta, por ejemplo, lectura o aritmética. En este enfoque existe el supuesto subyacente de que la aptitud que se mide es un atributo fijo e invariable de la persona. Basándose en este supuesto, se considera que el factor más crucial es la precisión de la medición en términos de una clasificación exacta de los alumnos, hasta el punto de que sólo se conservan para su empleo en los tests aquellos ítems que sirven para distinguir con más eficacia a unos alumnos de otros. Se descartan otros ítems que quizás no distinguen tan bien a unos alumnos de otros, independientemente del valor educativo del contenido de la pregunta. Se requiere que todas las preguntas del test evalúen la misma destreza o rasgo, y por ello se presupone que el rendimiento relativo del alumno en cada pregunta refleja una medida lineal de capacidad. Por lo tanto, se espera que existan estrechas relaciones estadísticas entre el rendimiento en cada pregunta y el rendimiento en el test en su conjunto. En general, estos tests se denominan tests objetivos estandarizados y se hacen bajo condiciones estrictamente controladas.

Además de la precisión de la medición, la equidad también se ha tenido muy en cuenta a la hora de desarrollar estos tests. Dado que los tests se corrigen automáticamente, no puede producirse injusticia alguna derivada de aplicar diferentes estándares de calificación. Se pretende principalmente medir la capacidad del alumno con independencia de sus antecedentes sociales o educativos, o de su etnia. Se llevan a cabo enormes esfuerzos para identificar y excluir cualquier ítem que suponga un sesgo, pudiendo favorecer o perjudicar a un subgrupo específico de alumnos. Las dificultades en torno a esta cuestión se comentan en la sección 2.7.

El objetivo inicial de medir la aptitud fue modificándose gradualmente al comprenderse cada vez más que en los tests educativos era imposible, y a menudo no deseable, separar la capacidad latente de los efectos de la experiencia educativa. También se fue comprendiendo cada vez más que las capacidades latentes, tales como la inteligencia, no constituyen atributos unidimensionales fijos del individuo. Medir la inteligencia de una persona no es lo mismo que, por ejemplo, medir su altura, proceso que quizás sí justifique que se siga afinando la precisión de la medición.

Los tests estandarizados se fueron utilizando cada vez más para medir los logros del alumno. Sin embargo, se conservó el enfoque eminentemente estadístico, al igual que el estilo de las preguntas. Los logros del alumno, en cuanto a los objetivos curriculares de la enseñanza en el aula, se inferían a partir de su rendimiento en un tipo muy restringido de tareas de evaluación. Esta evaluación de los logros del alumno en un sentido más amplio (por ejemplo, la de sus destrezas de expresión escrita) realizada por inferencia a partir de un tipo limitado de logro (las respuestas a preguntas de opción múltiple) sólo puede considerarse eficaz si se trata de una evaluación “de bajo riesgo”, y si ni alumnos ni profesores dan mucha importancia a la diferencia entre lo que debe enseñarse y lo que se está sometiendo a prueba.

Sin embargo, las pruebas psicométricas se han convertido en una evaluación “de alto riesgo” en algunos países, en los que están ligadas a la atribución de responsabilidades a colegios y profesores, y cuyos resultados influyen de forma significativa en las oportunidades que los alumnos pueden tener en la vida. En tales entornos, la naturaleza limitada de las destrezas y conocimientos evaluados tendrá inevitablemente una gran influencia en el modo en que los profesores enseñan en clase, influencia que se conoce como “efecto de repercusión”. Se ha determinado que el efecto negativo de esa enseñanza centrada en los exámenes constituye una desventaja importante de los tests objetivos estandarizados (por ejemplo, Hoffmann, 1962; Resnick y Resnick, 1992).

Se ha cuestionado la creencia de que los tests pueden y deben medir un sólo atributo unitario, incluso la inteligencia se considera generalmente en la actualidad como un concepto multidimensional (véase, por ejemplo, Gardner, 1983). Aparte de la restringida uniformidad de su enfoque, se considera especialmente desacertado que los tests estandarizados fomenten una visión del aprendizaje atomizada e inconexa. Sin embargo, es cierto que los tests estandarizados tienen ventajas importantes, como la de su rentabilidad en cuanto a tiempo y costos cuando se realizan a gran escala, y la de que sus resultados tienen un alto grado de fiabilidad.

La otra tradición, la evaluación del rendimiento, comprende una gran variedad de formas de evaluación en las que se espera que los alumnos realicen tareas que reflejen directamente los conocimientos y destrezas que han aprendido en clase. Por lo tanto, la evaluación del rendimiento incluye: resolución de problemas, redacciones, trabajos de proyecto y exámenes.<sup>1</sup> Debido a que las tareas de evaluación han de ser variadas, y a que las respuestas de los alumnos (o la demostración por su parte de una destreza concreta) han de ser substanciales, la evaluación del rendimiento consume mucho tiempo y dinero. Debe procurarse que las tareas de evaluación se diseñen de modo que reflejen las destrezas y conocimientos que se desea evaluar. La evaluación del rendimiento también requiere la opinión profesional de un examinador que califique el rendimiento y proporcione un resultado mensurable del mismo. La necesidad de dichas personas aumenta los costos y también reduce la fiabilidad de la evaluación, ya que inevitablemente existirán diferencias de opinión sobre el trabajo de los alumnos. Wood (1991, cap. 5), por ejemplo, comenta una serie de estudios que demuestran que la correlación entre las puntuaciones concedidas a una redacción de forma completamente independiente por dos examinadores rara vez supera el 0,6. Sin embargo, la ventaja principal de la evaluación del rendimiento es que está relacionada con los objetivos curriculares y apoya la enseñanza en clase, intentando así hacer un uso positivo de la enseñanza centrada en los exámenes (Wiggins, 1989). Una subcategoría de la evaluación del rendimiento, denominada evaluación auténtica, implica que los profesores evalúan las actividades de clase según se van realizando durante el transcurso de la enseñanza en clase. Normalmente, esta evaluación auténtica se considera parte de la evaluación formativa, pero también puede serlo del modelo de evaluación sumativa.

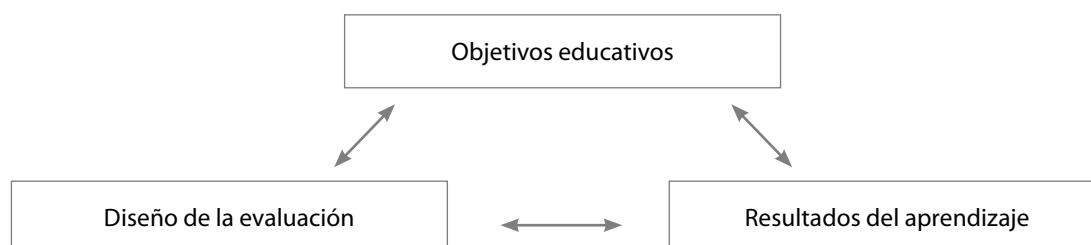
---

<sup>1</sup> Distintos autores han utilizado interpretaciones diferentes del término “evaluación del rendimiento” para adaptarlo a sus propios fines. En cuanto a esta publicación, se utiliza el término en su sentido más amplio, que incluye todas las formas de evaluación en las que se evalúa directamente el modo en que los alumnos demuestran que han alcanzado los objetivos de aprendizaje del curso.

## 2.4 Evaluación y aprendizaje

Los tests de logros objetivos estandarizados se desarrollaron sobre la base de la teoría conductista, según la cual, el aprendizaje es el resultado de establecer numerosas relaciones estímulo-respuesta (Resnick y Resnick, 1992); las destrezas y conocimientos complejos pueden descomponerse en bloques discretos y descontextualizados; y el aprendizaje se considera lineal y secuencial.

La evaluación del rendimiento, si se diseña y lleva a cabo adecuadamente, refleja una forma más moderna de entender la naturaleza del aprendizaje. La investigación en las áreas de la psicología cognitiva y el aprendizaje (véase, por ejemplo, Resnick, 1989; Shepard, 1991) revela que, ampliando sus estructuras de conocimiento previas (o esquema), es como mejor profundizan los alumnos en su comprensión a través de la interpretación y construcción de conocimientos y destrezas. El aprendizaje a menudo se produce de forma irregular, sin seguir una secuencia lógica desde lo sencillo a lo complejo. Este enfoque constructivista le reconoce al alumno un papel más activo, y también reconoce la importancia que tiene el contexto en la eficacia del aprendizaje (Murphy, 1999). Si la evaluación ha de apoyar la enseñanza y el aprendizaje eficaces, entonces debe diseñarse en torno a la moderna teoría constructivista del aprendizaje (Black, 1999; Shepard, 1992; Wood, 1998; Lambert y Lines, 2000). La evaluación formativa se vincula más directamente al modo en que los alumnos aprenden y, a veces, se la denomina evaluación *para* el aprendizaje, mientras que a la evaluación sumativa se la conoce como evaluación *del* aprendizaje. Se trata de una distinción engañosa que infravalora la importante influencia que tiene la evaluación sumativa en lo que de verdad se aprende en clase. Toda evaluación debe apoyar un aprendizaje apropiado. La evaluación sumativa no es sólo una actividad que se realiza después de haber tenido lugar el aprendizaje, sino que debe diseñarse para que su papel esté integrado en la enseñanza y el aprendizaje de la asignatura. Su nivel de integración debe ir más allá de lo que implica el término "efecto de repercusión", y puede expresarse en el paradigma siguiente (del paradigma de Furst, 1958, en Frith y Macintosh, 1984).



## 2.5 Evaluación normativa y evaluación por criterios

Los términos evaluación normativa y evaluación por criterios constituyen otra dimensión que puede diferenciar a los sistemas de evaluación, una dimensión que refleja los medios por los que se informa de los resultados de un proceso de evaluación basándose en una escala coherente a lo largo de un período de tiempo o de distintos sucesos. Uno de los métodos, generalmente asociados con los tests estandarizados, consiste en probar el test con una muestra típica de alumnos, y utilizar los resultados (a los que, por definición debe corresponder una distribución normal o curva campaniforme) como escala de referencia para la calificación de cualquier alumno que haga el mismo test después. Esto se denomina evaluación normativa, y el proceso de derivar una distribución estándar de las puntuaciones a partir de la prueba inicial se llama normalización. La evaluación normativa no implica necesariamente que se aplique una distribución fija a cada conjunto de resultados del test (la distribución fija sólo se utiliza en la normalización inicial). La distribución de las puntuaciones de posteriores alumnos puede ser distinta de la distribución normal. En principio, los tests normativos pueden utilizarse para informar de los cambios en el rendimiento del alumno a lo largo del tiempo; no informan únicamente sobre la puntuación de un alumno en comparación con sus compañeros de clase. A lo largo de los años, los resultados pueden variar en un sentido ascendente

o descendente con el uso reiterado de un determinado test. Cannell (1988), por ejemplo, descubrió que los alumnos que cursaban estudios primarios en los 50 estados de EE.UU. presentaban puntuaciones por encima de la media en tests normativos de destrezas básicas establecidos a nivel nacional. Se intentó encontrar distintas explicaciones al resultado de este descubrimiento, y la de que pudiera deberse a una mejora de la enseñanza y del aprendizaje fue una de las menos aceptadas.

A pesar del potencial que tienen los tests normativos para medir cambios en el estándar de logro del alumno a lo largo del tiempo, el término evaluación normativa también se aplica ahora habitualmente a aquellos sistemas de evaluación que imponen la misma distribución fija de resultados en reiteradas ocasiones. En estos sistemas, la calificación de cada alumno sólo mide en realidad su posición en el orden de clasificación de todos los alumnos evaluados en esa ocasión. Esto dice menos sobre lo que ha conseguido cada alumno que sobre quién lo ha hecho mejor o peor que dicho alumno, lo que naturalmente puede ser bastante adecuado para ciertos instrumentos de selección basados en cuotas.

Glaser (1963) fue el primero que propuso la noción de evaluación por criterios. Dicha evaluación supuso un cambio de dirección respecto a las pruebas de aptitud, dirigiendo el énfasis hacia la medición de los logros del alumno “con respecto a un ámbito de conducta bien definido” (Popham, 1978). También llevó a otro enfoque que tuvo influencia en EE.UU., llamado instrucción impulsada por la medición (Popham, 1987), que aprovecha el “efecto de repercusión” de las pruebas criterioles “de alto riesgo” para influir en el programa de instrucción que desemboca en la prueba. Sin embargo, debe señalarse que en su forma correcta, tal y como se definió inicialmente, la evaluación por criterios continúa basándose en la teoría conductista del aprendizaje y conlleva tests objetivos estandarizados. Los tests se construyen en torno al rendimiento del alumno en un ámbito restringido del aprendizaje (por ejemplo, la suma de fracciones simples) y suponen un conjunto de destrezas discretas y jerarquizadas. Las principales características que distinguen a los tests criterioles son las siguientes:

1. los ítems del test criterial son seleccionados de forma que representen unidades discretas de aprendizaje del alumno, y
2. el resultado del test depende de si el alumno ha alcanzado una puntuación de corte predeterminada teóricamente, en lugar de cuál sea su puntuación en comparación con una distribución predeterminada del rendimiento.

El resultado del test criterial tradicional es que se ha demostrado o que no se ha demostrado que se domina el ámbito pertinente. Los tests criterioles y los tests normativos difieren más en cuanto al análisis e interpretación de las respuestas del alumno que en cuanto al tipo de preguntas que contienen.

Fuera de EE.UU., los términos evaluación normativa y evaluación por criterios a menudo se utilizan sin mucha precisión para referirse a muy distintas filosofías sobre construcción de instrumentos de evaluación, información de notas obtenidas y medición de estándares de logro. Sin embargo, no debe sobreestimarse la diferencia entre los dos enfoques. La normalización inicial de los tests normativos puede establecer el estándar por el que se medirá a los futuros alumnos, y el estándar de logro que se requiere para demostrar que se domina la materia en un test criterial a menudo se decide con relación al rendimiento previsto para el conjunto del alumnado: las normas pueden representar estándares, y los criterios se establecen basándose en datos normativos.

Se ha intentado varias veces introducir la evaluación por criterios en el Reino Unido, en el contexto de los exámenes del GCSE (Certificado General de Educación Secundaria), introducidos en 1986 (Orr y Nuttall, 1983; SEC, 1984; Kingdon y Stobart, 1987). También se intentó introducir la evaluación por criterios en las pruebas del currículo nacional del Reino Unido (Brown, 1988). Dichos intentos fracasaron. La evaluación por criterios sólo es apropiada, en un sentido estricto, para tareas relativamente directas, fáciles de definir y uniformes; y no es apropiada para medir el rendimiento en áreas de materias más complejas. Para poder hacerlo se requiere, o bien un sistema de evaluación costoso y complejo, o criterios generales expresados de forma imprecisa en los que quepan variaciones del rendimiento de los alumnos en diferentes partes de la evaluación global, lo que no se adapta al modelo original. Un modelo de evaluación por criterios riguroso

exigiría el dominio de todos los aspectos de una asignatura antes de que pudiera decirse que el alumno había “aprobado”, e informaría de los resultados conseguidos por el alumno en sus peores aspectos. En el Reino Unido y otros países, se ha adoptado un enfoque más generalizado bajo diversas denominaciones, como por ejemplo, evaluación con referencia a estándares, (Sadler, 1987), evaluación basada en estándares, evaluación basada en resultados, evaluación con referencia a constructos (Black, 1998) o evaluación relativa a criterios.

## 2.6 Validez y fiabilidad: panorama general

Según la definición corriente, la validez de una evaluación reside en hasta qué punto mide realmente lo que se supone que ha de medir. El término fiabilidad se utiliza para definir la exactitud de la medición resultante de una evaluación y el grado de probabilidad de que se produzca el mismo resultado en circunstancias ligeramente distintas. Se considera que una evaluación es fiable si al repetir la evaluación en distintas ocasiones el alumno obtiene el mismo resultado, y también si la corrección la realizan distintos evaluadores. La validez y la fiabilidad se consideran en general características esenciales de todo sistema de evaluación, en especial de los “de alto riesgo” porque el resultado es de gran importancia para el alumno o para el profesor. Ambas características presentan de hecho muchas modalidades. Existen diferentes clases de validez y también de fiabilidad.

Los términos validez y fiabilidad, y las formas de abordar su medición, provienen fundamentalmente de la psicometría. El propio test define el constructo, o aptitud, que se mide, y la uniformidad se garantiza eliminando preguntas que produzcan respuestas erráticas por parte de quienes hacen el test. La construcción de tests estandarizados como grandes acumulaciones de ítems, que se comportan de forma similar y predecible, conduce casi inevitablemente a un alto grado de fiabilidad y a predicar con confianza su validez, porque puede demostrarse que todo el test se refiere a una sola área de aptitudes o destrezas, que puede etiquetarse fácil y adecuadamente. En las pruebas psicométricas, la fiabilidad y la validez (en sentido restringido) acaban estando inextricablemente entrelazadas. La validez consiste principalmente en determinar el único constructo que ha de medir un test dado; y la fiabilidad se refiere al grado de coherencia del comportamiento de los diferentes ítems del test, definido en términos de la correlación de las respuestas dadas por los alumnos a estos diferentes ítems.

En el contexto de la evaluación del rendimiento, los conceptos de validez y fiabilidad asumen interpretaciones ligeramente distintas. El de validez se amplía para abarcar las consecuencias sociales y educativas de realizar evaluaciones, reconociendo la interacción entre evaluación y enseñanza. Los instrumentos de evaluación del rendimiento rara vez se diseñan para medir un constructo uniforme y muy concreto y se acepta que, por ejemplo, las destrezas de comprensión y expresión escrita no se encuentren en la evaluación totalmente separadas de las destrezas de la asignatura. La interpretación de los resultados de la evaluación y los usos que se les da a los mismos se convierten en parte de la validez de la evaluación. Se pone menos énfasis en la precisión y exactitud de la medición. Se acepta que no es posible lograr niveles altos de fiabilidad técnica en un sistema de exámenes (por ejemplo, Wood, 1991; Satterley, 1994).

Sin embargo, dada la amplitud y complejidad de las destrezas educativas y de los logros que intenta normalmente abordar la evaluación del rendimiento, particularmente en lo que se refiere a alumnos que se encuentran al final de su educación secundaria, existen muchas dudas sobre el valor de un concepto como el de “calificación verdadera”, que los procesos de evaluación deban intentar determinar para cada alumno. Las diferentes áreas de destrezas y la amplia gama de posibles marcos contextuales en los que puede demostrarse destreza y conocimientos son tan diversas que imposibilita una medición precisa y completa.

Además, la capacidad del alumno en toda esta gama potencial de posibles logros no es una cantidad estática e invariable, como lo sería su altura, sino que es algo de naturaleza más dinámica y variable. La precisión de la medición llevada a cabo por cualquier instrumento de evaluación concreto, incluso si pudiera lograrse, no sería significativa por cuanto que representaría los logros del alumno solamente en ese grupo concreto de

tareas y en ese momento concreto. La variabilidad del alumno es un factor decisivo que hace imposible la existencia de un alto grado de fiabilidad. Debe aceptarse que sólo puede haber cierta aproximación, lo que no debe interpretarse como que las organizaciones responsables de las evaluaciones puedan permitirse hacer caso omiso de la exigencia ineludible de que los resultados que expidan en un nivel determinado deban ser lo más fiables posible.

Gipps (1994) ha defendido enérgicamente un cambio de paradigma a fin de redefinir el concepto de fiabilidad en el contexto de la evaluación educativa, en oposición a las pruebas psicométricas, afirmando (p.167):

“La evaluación no es una ciencia exacta, y debemos dejar de presentarla como tal. Esto es, naturalmente, parte de la posición del post-modernismo: una suspensión de la creencia en el estado absoluto del conocimiento ‘científico’. La postura modernista sugiere que es posible ser un observador desinteresado, mientras que la post-modernista mantiene que no es posible tal distanciamiento... El paradigma constructivista no acepta que la realidad sea fija e independiente del observador sino más bien que la realidad la construye el observador, por lo tanto, existen múltiples construcciones de la realidad. Este paradigma negaría entonces la existencia de algo como una ‘calificación verdadera’.”

Para resumir, la validez y la fiabilidad se definieron inicialmente de forma restringida, haciendo posible que se uniesen y se ajustasen cómodamente al marco teórico de las pruebas psicométricas. Capacidad y logro se consideran ambos como un atributo unificado del individuo, que puede medirse con exactitud y precisión. En el contexto más amplio de la evaluación del rendimiento, o de la evaluación educativa, la validez y la fiabilidad adoptan significados más amplios y se ha reconocido la existencia de un conflicto entre ellas (por ejemplo, Harlen, 1994). Las mejoras del grado de validez del constructo a menudo sólo pueden lograrse a costa de la fiabilidad, y viceversa. Una correcta evaluación sólo se consigue haciendo concesiones, y la naturaleza del equilibrio entre fiabilidad y validez dependerá del contexto y propósito de cada sistema de evaluación concreto. Gipps (1994) hace referencia al concepto de “confiabilidad”, entendido como la combinación óptima de validez y fiabilidad con relación a un determinado tipo de evaluación. En la evaluación formativa, se puede dar preponderancia a la validez, mientras que en la evaluación sumativa se debe prestar la misma atención a la validez que a la fiabilidad.

## 2.7 Sesgo

Puede definirse como la diferencia que se produce en el resultado de un proceso de evaluación y que no tiene que ver con una auténtica diferencia de la aptitud o de los logros objeto de medición. El sesgo puede surgir de los propios ítems del test/tareas de evaluación, o del proceso de corrección. En este último caso, el sesgo se convierte en una cuestión de fiabilidad de la corrección.

El sesgo que surge de las propias tareas de evaluación es un problema de principios más significativo. En la construcción de los tests psicométricos, cuando se demuestra que un ítem produce características de respuesta inusuales durante las pruebas previas, o muestra características de respuesta sustancialmente distintas en diferentes subgrupos del alumnado (“Funcionamiento Diferencial del Ítem”/“Differential Item Functioning” o DIF) se considera sesgado y se elimina del test. Los subgrupos de alumnos pueden definirse por género, etnia, clase social o competencia lingüística, de hecho, por cualquier característica definitoria que pueda considerarse no pertinente con respecto al constructo objeto del test. No obstante, las pretensiones de la existencia de sesgo a favor o en contra de subgrupos de alumnos concretos no siempre tienen una justificación evidente. En los años iniciales del desarrollo de los tests de inteligencia, se excluyeron aquellos ítems que daban lugar a una diferencia significativa de respuesta entre ambos géneros. Esto se basaba en la suposición de que no debía existir diferencia en el constructo de inteligencia entre varones y mujeres, y por tanto los ítems que revelasen esa diferencia debían estar midiendo algo sin pertinencia para el caso. Esa opinión es, cuando menos, debatible, y varios autores han ofrecido explicaciones para las diferencias en las mediciones de inteligencia entre diferentes grupos de personas, relacionadas con factores biológicos, del

entorno, o socioeconómicos, así como con la naturaleza de los propios tests. En el desarrollo de las pruebas de inteligencia se ha puesto de manifiesto mayor preocupación por la fiabilidad de la medición que por la naturaleza de lo que de hecho se está midiendo, lo cual se ha moldeado para adaptarse a las exigencias de un alto grado de fiabilidad. Pueden existir aspectos importantes de un constructo que estén legítimamente ligados a ciertas características de grupos dentro del conjunto del alumnado. Las implicaciones de estas diferencias para la enseñanza y el aprendizaje tienen un gran potencial. Un enfoque más adecuado es incluir en los tests una gama equilibrada de ítems que dan lugar a un rendimiento diferencial por parte de los diferentes subgrupos del alumnado, de modo que ningún subgrupo se encuentre en desventaja. Cualquiera que sea el enfoque que se adopte, puede establecer limitaciones considerables a la hora de diseñar un test.

Las decisiones sobre el sesgo o legitimidad de ciertos ítems de un test deben basarse en el modo en que cada ítem pueda ser vinculado explícitamente al constructo subyacente, y en cuáles puedan ser los posibles factores que introduzcan el sesgo, en lugar de en motivos puramente estadísticos basados en la calibración del ítem. Goldstein (1996a) y Humphreys (1986) han sugerido que es conveniente distinguir entre “diferencia”, que es un hecho determinado objetivamente, y “sesgo”, que es un juicio sobre la pertinencia de la diferencia. Black (1998, p. 50) propone las siguientes seis posibilidades como aquéllas que más a menudo pueden hacer que las preguntas sean injustas en cuanto a su impacto en alumnos distintos.

- El contexto en el que se plantea la pregunta (por ejemplo, los juguetes mecánicos y ciertos deportes favorecen a los niños, las muñecas y el trabajo doméstico favorecen a las niñas).
- Las preguntas consistentes en redacciones sobre temas impersonales favorecen a los niños, mientras que las que versan sobre relaciones humanas favorecen a las niñas.
- Las preguntas de opción múltiple favorecen a los niños.
- El trabajo de clase o de proyecto favorece a las niñas.
- Algunas preguntas pueden ser inteligibles solamente en el seno de ciertas culturas, por ejemplo, una pregunta sobre ancianos que vivan solos puede resultar bastante ajena a ciertas culturas, o una pregunta que trate de un papel típicamente masculino o femenino en una cultura puede aparecer muy fuera de lugar en otra.
- Una pregunta que utilice lenguaje o convenciones de una clase social favorecerá a los alumnos de dicha clase.

Se ha realizado una extensa investigación de estas diferencias, que se encuentran ya consolidadas (véase, por ejemplo, Wood, 1991, cap. 14). El modo en que deben responder ante tales diferencias quienes diseñan las evaluaciones no está siempre tan claro. Los instrumentos de evaluación deben diseñarse de modo que, mediante diversos tipos de tareas y preguntas, se reduzcan las consecuencias generales del sesgo. Deben evitarse todos los estereotipos culturales o de género (explícitos o no). El contenido de las preguntas debe examinarse minuciosamente para evitar que pertenezcan a alguna de las categorías que se sabe que introducen factores de injusticia, y la realización de pruebas previas de las preguntas en muestras de los diferentes subgrupos del alumnado pueden revelar casos ocultos de ello. No obstante, si se excluyen todos los tipos de preguntas con sesgo y todos los marcos hipotéticos, quedan pocas opciones disponibles para diseñar evaluaciones y construir preguntas, y las limitaciones resultantes tendrán un efecto negativo en la validez de la evaluación. Aparte de evitar los peligros obvios e innecesarios, la solución más razonable parece ser la de un enfoque del diseño de la evaluación equilibrado, utilizando distintos tipos de tareas y formatos de evaluación.

También causa preocupación saber cuántos subgrupos diferenciados del alumnado deben tenerse en cuenta. ¿Debe tenerse en cuenta a los alumnos con diferentes estilos de aprendizaje, o a los no aptos temperamentalmente para los tests o exámenes formales? Según han afirmado Hieronymus y Hoover (1986), si se considera que las diferencias de interés y motivación son factores de sesgo, puede decirse que todas las tareas o métodos de evaluación tienen un cierto grado de sesgo. Por ejemplo, resulta inevitable que los pasajes de texto utilizados en los exámenes de lengua tengan mayor interés para unos alumnos

que para otros. Al final, la preocupación sobre la posibilidad de sesgo en diferentes tipos de preguntas y contextos a menudo se reduce a una cuestión sociopolítica.

La equidad en la evaluación, que incluye evitar el sesgo, es un tema de suma importancia, especialmente en ciertos países donde cualquier sesgo demostrable en un instrumento de evaluación puede incluso desembocar en un litigio. Sin embargo, demostrar que se trata de sesgo, y no de diferencia en el rendimiento, a menudo requiere una valoración muy precisa, estrechamente ligada al contexto social concreto en el que se lleva a cabo la evaluación. Gipps y Murphy (1994) concluyeron su libro *A Fair Test? Assessment, Achievement and Equity* (¿Un test justo? Evaluación, logro y equidad) diciendo: "No existe ni puede existir ningún test que podamos calificar de justo. La situación es demasiado compleja y la noción demasiado simple". Sin embargo, eso no significa que quienes diseñan las evaluaciones y redactan las preguntas no deban hacer todo lo que esté en su mano para reducir el efecto del sesgo y de la injusticia. Gipps y Murphy también defienden que quienes diseñan las evaluaciones deben tener como objetivo la igualdad de oportunidades y de acceso a la evaluación, en lugar de la igualdad de resultados que se consigue manipulando ítems de los tests según las estadísticas de las respuestas. Cuestionan hasta qué punto se justificaría, por ejemplo, incluir pruebas de opción múltiple en los exámenes de lengua para mejorar el rendimiento relativo de los varones, ya que distorsionaría la validez de la evaluación según nuestra concepción de definición de la materia.

Se acepta, en general, que la falta de justicia en el proceso de evaluación es sólo uno de los factores que contribuyen a la falta de equidad en la educación, y posiblemente uno de los menos significativos. El rendimiento diferencial en un test por parte de diferentes subgrupos puede ser el resultado de factores que tengan bastante poco que ver con el propio test. Hay muchos otros factores causantes de falta de equidad en la educación que tienen un impacto decisivo en los logros del alumno, por ejemplo, las diferencias de calidad de enseñanza en un mismo colegio, las diferencias de nivel de recursos de diferentes colegios y en diferentes áreas geográficas, y las diferencias en las circunstancias sociales y en el nivel de apoyo familiar que recibe cada alumno. Cualquiera de estos factores, o todos ellos, pueden influir de forma significativa en las perspectivas individuales de éxito educativo de un modo tal que ningún proceso de evaluación, por justo que fuese, podría compensar. Smith y Tomlinson (1989), por ejemplo, descubrieron que la eficacia del colegio era un factor mucho más importante para la existencia de diferencias en los resultados de los exámenes que la etnia de los alumnos, señalando que los intentos de ajustar los instrumentos de evaluación para remediar diferencias en el rendimiento por parte de diferentes grupos étnicos pueden ser a veces inadecuados.

Las consideraciones de esta naturaleza constituyeron la base lógica para que se realizasen pruebas de aptitud en lugar de pruebas de logro, pero se ha llegado a comprender que no es posible evaluar la aptitud, capacidad o potencial puros, separados del origen social y de la experiencia educativa. Tampoco es posible juzgar el logro educativo de manera objetiva independientemente del contexto social y la cultura. El concepto de éxito en la educación se define y mide en cualquier sociedad según los estándares de un pequeño sector de la misma.

Otro aspecto del sesgo que debe combatirse es la posibilidad de que una tarea de evaluación discrimine negativa e injustamente a los alumnos con necesidades educativas especiales tales como dislexia, síndrome de déficit de atención, o problemas de visión. Las condiciones de realización de las tareas de evaluación deben hacer las concesiones apropiadas para estos alumnos, de modo que puedan demostrar su nivel de logro educativo en condiciones de igualdad con respecto a los demás alumnos.

El sesgo que se produce al corregir las pruebas puede tener varios motivos, tales como la actitud personal frente a la pulcritud de la escritura del alumno (por ejemplo, Hughes et ál., 1983), el trato preferente en función del género del alumno (en caso de que pueda saberlo o sospecharlo el evaluador), y la indebida atención que se preste a factores como el formateado, los signos de puntuación y la ortografía, que pueden no tener una pertinencia significativa en ciertos contextos de evaluación. Para resolver estas cuestiones, los examinadores deben recibir formación adecuada y su trabajo ha de ser revisado.

## La evaluación en el Programa del Diploma: objetivos y enfoques

En esta sección se describe el enfoque filosófico adoptado en el contexto de la evaluación formal que se utiliza en el Programa del Diploma del BI. Esto incluye el modo en que se incorporan los principios básicos de la evaluación (descritos en la sección 2) y proporciona la base sobre la que se construyen las estructuras y los procesos (véanse las secciones 4 y 5). Por su misma naturaleza, la evaluación formal del Programa del Diploma es una evaluación sumativa, diseñada para registrar los logros alcanzados por el alumno al finalizar el curso o hacia el final del mismo. Debe señalarse, no obstante, que muchos de los instrumentos de evaluación también pueden utilizarse formativamente durante la enseñanza y el aprendizaje, y se anima a los profesores a que así lo hagan. Esto se aplica en particular a las tareas de evaluación interna (véanse las secciones 4.2 y 5.3).

El sistema de evaluación del Programa del Diploma es una evaluación del rendimiento, “de alto riesgo” y por criterios. Se basa en los siguientes objetivos, los cuales desarrollaremos a lo largo de esta sección.

1. La evaluación debe apoyar los objetivos curriculares y filosóficos del programa, fomentando las buenas prácticas en clase y el aprendizaje adecuado del alumno.
2. Los resultados publicados de la evaluación (es decir, las calificaciones finales de las asignaturas) deben tener un nivel lo suficientemente alto de fiabilidad, adecuado a una titulación “de alto riesgo” de acceso a la universidad.
3. La evaluación debe reflejar la conciencia internacional del programa en la medida de lo posible, debe evitar el sesgo cultural y debe hacer las concesiones adecuadas para los alumnos que estén trabajando en su segunda lengua.
4. La evaluación debe prestar la adecuada atención a las capacidades cognitivas superiores (síntesis, reflexión, evaluación, pensamiento crítico), así como a las más fundamentales (conocimiento, comprensión y aplicación).
5. La evaluación de cada asignatura debe incluir una gama apropiada de tareas e instrumentos/ componentes que garantice que evalúen todos los objetivos específicos de la asignatura.
6. Para evaluar los logros alcanzados por el alumno y determinar las calificaciones finales en las distintas asignaturas debe seguirse principalmente la opinión profesional de examinadores supervisores experimentados, apoyada en información estadística.

### 3.1 Apoyo a los objetivos curriculares

Aunque la lista anterior no se presenta en orden descendente de prioridades, y algunos de los objetivos generales están interrelacionados, está claro que el objetivo general más importante del sistema de evaluación del Programa del Diploma es que debe apoyar y fomentar el aprendizaje adecuado del alumno. Ésta es la característica más valorada por quienes van a utilizar de algún modo el título del Diploma, principalmente instituciones de enseñanza superior (IBO, 2003a), y por los propios colegios y alumnos. La fiabilidad absoluta de los resultados de la evaluación, aunque enormemente importante en sí misma, no puede ser más prioritaria que el aprendizaje del alumno.

Existe un auténtico conflicto en el diseño de la evaluación entre las técnicas que pueden ofrecer las mediciones más precisas y fiables de ciertos aspectos de los logros del alumno, y las que miden y fomentan los logros educativos más deseables de los alumnos. Alec Peterson (1971) identificó muy bien este dilema, describiendo el desarrollo inicial de la evaluación en el Programa del Diploma de la siguiente manera:

“Lo que se necesita es un proceso de evaluación que sea lo más válido posible, en el sentido de que realmente evalúe las capacidades y la personalidad del alumno con relación a la próxima etapa de su vida, pero [que sea] a la vez lo suficientemente fiable como para garantizar a alumnos, padres, profesores e instituciones receptoras que se está haciendo justicia. Sin embargo, dicho proceso no debe, por su “efecto de repercusión”, distorsionar la buena enseñanza ni ser demasiado lento, ni absorber demasiados de nuestros escasos recursos educativos.”

Debe aprovecharse el fuerte impacto que tiene la evaluación “de alto riesgo” en la enseñanza y el aprendizaje, diseñando instrumentos de evaluación que fomenten una buena pedagogía y la implicación constructiva de los alumnos en su propio aprendizaje, teniendo en cuenta a la vez las líneas de pensamiento más recientes sobre teoría del aprendizaje (por ejemplo, Murphy, 1999). Si el Programa del Diploma tiene como meta formar jóvenes que sean “solidarios, informados y ávidos de conocimiento” y que lleguen a “adoptar una actitud activa de aprendizaje durante toda su vida y a ser compasivos” (Declaración de principios de IBO), entonces estas características deben verse reflejadas en el sistema de evaluación. Es un hecho ineludible que lo que no es objeto de evaluación no se valora tanto, y puede incluso llegar a pasarse por alto. Las aspiraciones expresadas en la declaración de principios deben ser apoyadas por el sistema de evaluación.

Las características personales que se desea que tengan los alumnos, expresadas en la declaración de principios de IBO, se ajustan muy bien a la teoría constructivista del aprendizaje, según la cual los alumnos se implican activamente en el proceso de aprendizaje, se responsabilizan de su propio aprendizaje y amplían sus conocimientos, comprensión y destrezas a través de la investigación. En los requisitos de la evaluación de varias asignaturas, se espera que el alumno muestre empatía respecto a perspectivas culturales distintas de la suya propia. Las cualidades de naturaleza afectiva, como son la solidaridad y la compasión, resultan más difíciles de incluir en la evaluación formal, pero deben, no obstante, verse representadas en el sistema de evaluación global. Esto se logra en gran medida a través del requisito de Creatividad, Acción y Servicio (CAS), aunque existen varias referencias a las prácticas de trabajo éticas en el sistema de evaluación.

En lo que se refiere a los principios de evaluación, la evaluación en el Programa del Diploma pone mucho énfasis en la validez consecuencial (véase el apéndice A.1), consciente de que el modo en que se lleve a cabo la evaluación tendrá una influencia decisiva en la forma en que se enseñe el Programa del Diploma en los colegios. Esta influencia se ha acentuado deliberadamente en los últimos años, proporcionando una cantidad cada vez mayor de comentarios a colegios y profesores sobre el rendimiento de sus alumnos en las evaluaciones y sobre distintas formas de mejorar dicho rendimiento.

Todavía deben llevarse a cabo estudios de investigación extensos sobre la validez predictiva de los resultados del Programa del Diploma, pero el pequeño número de estudios informales realizados y la cantidad sustancial de datos anecdóticos sugieren que la validez predictiva de los resultados del Diploma es alta. El modelo de evaluación (conjunto de instrumentos de evaluación) aplicado a cada asignatura se diseña para que tenga una base amplia, que abarque diversos tipos de datos, tanto para garantizar la validez de constructo como para mejorar la generalizabilidad de los resultados en la medida de lo posible.

## 3.2 Fiabilidad de los resultados

Aunque la fiabilidad de los resultados al nivel de calificaciones finales de las asignaturas debe ser prioritaria para un sistema de evaluación “de alto riesgo”, la precisión absoluta de la medición, dentro de un margen máximo de un punto en cada tarea emprendida por el alumno, no es posible ni incluso necesaria. La meta es tener al menos un 95% de confianza en que la calificación final de la asignatura sea “correcta”. En este sentido

se considera un resultado correcto el que se confirmaría si el trabajo lo vuelve a corregir un examinador con más experiencia. Se trata de un objetivo razonable para un sistema que depende mucho más de criterios cualitativos que de mediciones técnicas, y que utiliza sistemas de control de calidad para garantizar que se alcance este objetivo. El modelo de evaluación que se utiliza para generar un resultado en una asignatura comprende diversas tareas realizadas en distintos contextos en diferentes ocasiones. Esto contribuye a reducir la amenaza que supone para la fiabilidad una única tarea de evaluación realizada en un contexto en particular. No se consideran adecuadas las medidas de fiabilidad de la coherencia interna (véase el apéndice A.2) porque cada componente (instrumento de evaluación) puede contener deliberadamente formas de tarea variadas, o a veces un pequeño número de tareas. La fiabilidad de las formas paralelas (véase el apéndice A.2), con relación a los nuevos exámenes creados para cada convocatoria de evaluación, no es esencial al nivel de las puntuaciones que se conceden a los alumnos. Se acepta que pueda ser ligeramente más difícil o más fácil conseguir puntos en diferentes casos con el “mismo” instrumento de evaluación. La carga de la fiabilidad recae en la determinación de calificaciones que representen de modo coherente el mismo estándar de logro (véase la sección 5.6). Se pone mucho énfasis en garantizar la fiabilidad de la corrección mediante la utilización de esquemas de calificación detallados, criterios de evaluación y procedimientos de moderación (véanse secciones 5.4 y 5.5), y en reducir el sesgo del examinador. La fiabilidad de la corrección, en lo que afecta a la calificación final de la asignatura, también se mejora utilizando distintos examinadores para calificar partes diferentes del trabajo del alumno en una asignatura.

Se procura por todos los medios garantizar la fiabilidad de la calificación final, aplicando estándares coherentes apoyados en datos estadísticos al determinar las bandas de calificación. Se documentan y ejemplifican los estándares de calificación, y las decisiones que se toman sobre las bandas de calificación se comprueban por medio de varios indicadores estadísticos.

En resumen, según se avanza en el registro de los logros del alumno a través de los diferentes niveles (puntuación del componente, puntuación de la asignatura y calificación final de la asignatura) la fiabilidad de los informes aumenta. Se alcanza un alto grado de fiabilidad al nivel de la calificación final de la asignatura.

### 3.3 Dimensiones internacional e intercultural

El Programa del Diploma se estudia en más de 100 países por alumnos pertenecientes a incluso más nacionalidades. Además de alcanzar los objetivos académicos del programa, IBO pretende que, siguiendo el programa, los alumnos se conviertan en “jóvenes solidarios [...] capaces de contribuir a crear un mundo mejor y más pacífico, en el marco del entendimiento mutuo y el respeto intercultural”, y capaces de “entender que otras personas, con sus diferencias, también pueden estar en lo cierto” (Declaración de principios de IBO). La enseñanza tiene, por tanto, un contexto internacional así como una finalidad intercultural, que deben reflejarse en la evaluación. La lengua es un factor fundamental. Las evaluaciones del Programa del Diploma se llevan a cabo en inglés, francés y español. Los exámenes se preparan normalmente en inglés y se traducen al francés y al español. En la traducción se hace todo lo posible para lograr que el nivel de exigencia sea el mismo en las tres lenguas, dando lugar, si es necesario, a correcciones de la versión original de los exámenes en inglés. Si, al juzgar posteriormente el rendimiento real de los alumnos, parece que sin intención se ha dado una ligera ventaja o desventaja a quienes utilicen una lengua en particular, se realiza el correspondiente ajuste de las puntuaciones. Muchos de los examinadores supervisores son bilingües o trilingües.

El Programa del Diploma ofrece una amplia gama de cursos de segunda lengua para diferentes niveles de competencia y, además, garantiza que todos los alumnos puedan hacer un curso de literatura en la lengua en la que sean más competentes, siempre que en dicha lengua exista una colección de obras literarias suficientemente amplia como para formar una base adecuada de estudio.

Aparte de la cuestión de la lengua, muchas asignaturas del Programa del Diploma, y la evaluación de éstas, incluyen una importante dimensión intercultural. He aquí algunos ejemplos de ello:

- En el curso de literatura (Lengua A1), los alumnos deben estudiar algunas obras que fueron escritas originalmente en una lengua distinta de la que se estudia en el curso. Los alumnos deben redactar trabajos escritos sobre estas obras, que incluyan una perspectiva intercultural.
- En los cursos de lengua extranjera (por ejemplo, Lengua B), la lengua debe estudiarse en un sólido marco cultural y práctico. En la evaluación del uso de la lengua se tiene en cuenta el conocimiento del marco cultural.
- El curso de Historia incluye una sección obligatoria sobre historia mundial y desarrolla una perspectiva internacional para la explicación de la historia.
- En el curso de Economía, una parte sustancial del contenido de la asignatura versa sobre economía internacional y economía del desarrollo, esperándose que los alumnos entiendan la teoría y aplicación de la economía desde diferentes perspectivas nacionales y culturales.
- En el curso de Música, los alumnos deben llevar a cabo una investigación de la relación entre dos géneros musicales de diferentes culturas.

En algunas otras asignaturas, la cuestión de la diversidad cultural se aborda a través de la tolerancia de diferentes énfasis culturales que permite la estructura del contenido de la asignatura. Pueden encontrarse ejemplos de este enfoque en Biología, Química, Psicología y Artes Visuales. En las tres primeras, las estructuras de opciones de cada asignatura permiten que los colegios elijan hasta cierto punto el contenido del curso para adaptarlo a las formas culturales tradicionales de impartir la asignatura.

Los cursos del Programa del Diploma son, por tanto, tolerantes con respecto a las variantes culturales, a la vez que fomentan la tolerancia cultural. Esto plantea problemas de evaluación para mantener la equiparabilidad entre los enfoques opcionales que se permiten para parte de muchas asignaturas. La equiparabilidad de las evaluaciones de una asignatura siempre se ve comprometida cuando se pueden elegir las preguntas, cuando hay opciones, o cuando hay tareas de evaluación muy abiertas. Sin embargo, la creciente utilidad y la validez de la evaluación que estas estructuras ofrecen a alumnos en distintos marcos culturales en diferentes partes del mundo, restan importancia a tales preocupaciones. El factor crucial es que los alumnos sigan cursos adecuados a sus propios contextos culturales. El mismo curso de Biología, por ejemplo, no sería apropiado para todos. Naturalmente, se debe intentar conseguir un grado de equiparabilidad lo más alto posible. Puede obtenerse información sobre equiparabilidad analizando el rendimiento del alumno en aspectos troncales de la evaluación con relación a su rendimiento en aspectos optativos de ésta.

El interculturalismo es algo más que el conocimiento y comprensión de otras culturas. La actitud y la acción también son atributos importantes. Es difícil evaluar las actitudes a través de la evaluación normal de los colegios, que se centra más en el logro que en los atributos afectivos. No hay intención de incluir ningún tipo de trazado de perfiles psicológicos en el sistema de evaluación del Programa del Diploma. En lugar de ello, lo que los alumnos aportan mediante sus acciones, reflejando sus valores y actitudes, constituye parte del requisito de Creatividad, Acción y Servicio (CAS) (véase la sección 4.1). No hay una escala de logros o de calificaciones correspondiente a CAS, pero los colegios deben autenticar la participación satisfactoria de cada alumno del Programa del Diploma. Sin esta autenticación, no puede concederse el diploma al alumno. CAS tiene, por tanto, un efecto importante en el resultado global de la evaluación. El interculturalismo práctico se fomenta en el trabajo evaluado correspondiente a las asignaturas de Artes, y también pueden surgir posibilidades de involucrarse de manera práctica en el trabajo de evaluación interna correspondiente a algunas otras asignaturas (véase la sección 4.2).

La evaluación que se lleva a cabo en un contexto internacional plantea más retos en cuanto a la equidad de los que normalmente se encuentran en el seno de un sistema nacional. Preguntas que pueden ser perfectamente adecuadas en un marco nacional se convierten en inadecuadas en otro. Las preguntas que se refieren a deportes, viajes, entretenimiento, acontecimientos históricos, incluso al tiempo, deben prepararse con muchísimo cuidado. Puede parecer que el único modo de evitar este problema sea preparar

preguntas de examen carentes de todo contexto sociocultural excepto el mínimo común denominador. Sin embargo, hacerlo así no sólo limitaría mucho las preguntas de exámenes y las haría aburridas, sino que iría en contra de toda la filosofía de evaluación del Programa del Diploma y contra el correcto ejercicio de la actividad de evaluación en lo concerniente a garantizar su validez mediante tareas basadas en contextos. La evaluación y el trabajo contextualizados son esenciales para el buen aprendizaje. Existen dos posibles soluciones para este dilema. En primer lugar, puede proporcionarse a los alumnos información contextual complementaria, especificándose en el contenido del programa de la asignatura, mediante ejemplos en los que se basan las preguntas, o incluso en la misma pregunta de examen (siempre que no ocupe demasiada extensión y distraiga así de la finalidad de la evaluación). Un segundo método consiste en utilizar preguntas y tareas de evaluación más abiertas que permitan a los alumnos elegir el contexto en el que responder. En este último enfoque, la puntuación ha de centrarse en niveles de comprensión más profundos, más que en conocimientos directos del contenido de la asignatura, ya que no habrá base común de contenido. Esto está muy en la línea de la filosofía de evaluación del Programa del Diploma (véase la sección 3.4).

Incluso aplicando estos dos métodos, los alumnos pueden encontrarse frente a tareas de evaluación cuyos contextos no les resulten familiares con referencia a sus propios antecedentes socioculturales. También esto está en la línea del Programa del Diploma y de su filosofía de evaluación, dado que uno de los objetivos del programa es abrir las mentes de los alumnos a otras formas de hacer las cosas, aumentar su conciencia global y su competencia para operar en un entorno cultural que no les resulte familiar. Parte del requisito del pensamiento abstracto es que los alumnos deben poder aplicar sus conocimientos en situaciones que no les resulten familiares. Es apropiado incluir tales elementos en la evaluación, siempre que afecten por igual a alumnos de diferentes antecedentes culturales.

Se ha hecho referencia antes en esta publicación (véase la sección 2.7) a investigaciones que señalaban las cuestiones de equidad existentes en torno a los diferentes formatos de evaluación. Gran parte de la investigación en esta área se ha referido al género (Wood, 1991, cap. 14; Gipps y Murphy, 1994). En un entorno internacional, los diferentes formatos de evaluación también pueden provocar una falta de equidad con base cultural, relacionada con la experiencia educativa anterior. Puede que algunos alumnos del Programa del Diploma se hayan visto expuestos anteriormente a formas de evaluación que consistían sólo en tests de opción múltiple, o preguntas de desarrollo, o entrevistas orales, o portafolios o carpetas. Ésta es una razón más para incluir una gama de formatos lo más amplia posible en el sistema de evaluación del Programa del Diploma (véase la sección 3.5), de forma que todos los alumnos y profesores encuentren algunos formatos que les resulten familiares y más cómodos, y otros que les resulten menos familiares.

Otras dos cuestiones de equidad cultural en el sistema de evaluación del Programa del Diploma se refieren a los exámenes que se hacen en una segunda lengua y al papel del trabajo en grupo.

Un número significativo de alumnos del Programa del Diploma se presentan a exámenes en una lengua que no es aquella en la que son más competentes. En casi todos estos casos se trata del inglés, porque los alumnos que estudian en francés o español (las otras dos lenguas en las que se realiza la evaluación) suelen ser hablantes nativos. Debe procurarse de manera especial que la redacción de las preguntas no sitúe en desventaja a los hablantes de inglés como segunda lengua. Las frases deben ser cortas y, siempre que sea posible, las estructuras de las frases y su redacción deben ser sencillas. Sin embargo, no debe evitarse utilizar la terminología específica de la asignatura. Además, debe mostrarse tolerancia respecto a los errores de ortografía y gramática cuando se lleva a cabo la puntuación, excepto en los exámenes de lengua. Siempre que quede claro el significado y la comunicación no se obstaculice, no debe penalizarse este aspecto y el alumno debe tener la posibilidad de alcanzar la puntuación máxima.

Finalmente, el sistema de evaluación del Programa del Diploma, al igual que la gran mayoría de sistemas de evaluación formales, es muy individualista. Como señaló Brown (2002), esto se debe en gran medida a que el Programa del Diploma pertenece a la tradición europea occidental, y las sociedades europeas occidentales son individualistas por naturaleza. Se evalúa a los alumnos casi exclusivamente sobre lo que logran por sí mismos. Puede decirse que culturalmente esto carece de equidad, ya que hay algunas culturas en las que lo que aporta el individuo está siempre subordinado a lo que aporta un grupo más amplio: lo

que importa es lo que logra el grupo. También sucede que en términos de equidad individual, hay algunas personas que trabajan mejor en equipo que individualmente, y viceversa. Además, es práctica habitual que, tanto en el aula como en el mundo laboral, los individuos trabajen de forma interdependiente más que independiente.

El trabajo en grupo plantea problemas importantes a la hora de realizar la evaluación en cuanto a identificar de forma fiable quién ha aportado qué y quién se ha beneficiado (o perjudicado) injustamente por el trabajo de otros. Sin embargo, en interés de la validez de la evaluación, así como de la equidad cultural, debe hacerse más por incluir el trabajo de grupo en muchos sistemas de evaluación. Efectivamente, el Programa del Diploma incluye un elemento de trabajo de grupo. En todos los cursos de Ciencias Experimentales, los alumnos deben participar en un proyecto interdisciplinario, que por naturaleza requiere trabajo de grupo. Uno de los criterios de evaluación aplicados al trabajo práctico en las ciencias valora si el alumno trabaja bien en equipo, y el proyecto interdisciplinario es un contexto apropiado en el que los profesores pueden evaluar esto.

En un sentido ligeramente distinto, el curso de Música tiene un componente de interpretación, en el que se puede optar por la interpretación en grupo. En este caso, a cada alumno del grupo se le concede la misma puntuación, de acuerdo con la interpretación de todo el grupo. Corresponde a cada miembro del grupo intentar asegurarse de que la interpretación del grupo en su conjunto sea la mejor posible. En el curso de Teoría del Conocimiento, los alumnos tienen que hacer una exposición oral sobre un tema ante el resto de la clase. Esta exposición debe formar parte integral de la enseñanza del curso. Las presentaciones pueden ser individuales o en grupo, pero si se hacen en grupo, el profesor concede los puntos a cada alumno según su aportación individual.

Estos ejemplos forman sólo una pequeña parte de la evaluación global del Programa del Diploma, y es cierto que el trabajo de grupo carece todavía de suficiente representatividad en la estructura global.

## 3.4 Capacidades cognitivas superiores

A menudo se afirma que vivimos en la sociedad del conocimiento. No ha de entenderse que esto signifique que la adquisición y retención de información fáctica sean de la máxima importancia. La explosión de información en los últimos tiempos hace imposible que los individuos logren alcanzar un dominio de muchas áreas del conocimiento. Las destrezas o capacidades académicas más valoradas hoy en día son las de acceder, ordenar, filtrar, sintetizar y evaluar información, y construir creativamente el conocimiento. Según aumenta la tasa de desarrollo y cambio en muchas sociedades, aprender a aprender se convierte en una capacidad o destreza más valiosa que la de simplemente aprender conocimientos y conceptos. Peterson (2003), la persona que más que ninguna otra ha configurado la filosofía educativa de IBO, ha dicho aproximadamente lo mismo, al afirmar que "lo que importa no es la absorción y regurgitación de hechos o de interpretaciones predigeridas de hechos, sino el desarrollo de las facultades de la mente o de formas de pensar que pueden aplicarse a nuevas situaciones y a la presentación de nuevos hechos según van surgiendo".

Si la evaluación ha de conservar su validez, las anteriores consideraciones deben influir en ella de forma fundamental. No basta ya con una evaluación que se refiera sólo a recordar conocimientos, conceptos y técnicas rutinarias. Si las capacidades que se espera que los alumnos de hoy desarrollen están cambiando, o mejor dicho, ampliándose para abarcar una mayor diversidad, entonces los instrumentos de evaluación deben hacer lo mismo. El sistema de evaluación del Programa del Diploma pretende deliberadamente prestar mucha atención a las llamadas capacidades cognitivas "superiores" (Bloom et ál., 1956). Puede que exista desacuerdo sobre la naturaleza jerárquica de los niveles que propone Bloom o sobre el número de niveles, pero su taxonomía de objetivos educativos continúa proporcionando un marco útil a través del cual expresar la diversidad de capacidades que se requiere. Las capacidades cognitivas superiores a las que se refiere Bloom exigen ciertamente que se utilice un tipo distinto de evaluación. Las capacidades de

análisis, síntesis y evaluación del alumno sólo pueden medirse adecuadamente exigiéndole que analice, sintetice y evalúe con cierta extensión. La evaluación del rendimiento es la única forma realista de evaluar los logros del alumno en estas áreas, y dado que los resultados de tal actividad no pueden ser prescritos ajustadamente, tales evaluaciones deben ser relativamente desestructuradas y abiertas. Inevitablemente esto suscita más preocupaciones sobre la fiabilidad de la corrección, al poder existir muchas respuestas distintas pero correctas.

Sin embargo, no puede hacerse caso omiso de las exigencias de validez de constructo cuando éste incluye capacidades productivas sofisticadas y complejas. Uno de los objetivos generales más antiguos del Programa del Diploma ha sido formar alumnos que sean pensadores críticos (Hill, 2002). Los objetivos específicos de las asignaturas del programa incluyen necesariamente una significativa representación de las capacidades cognitivas superiores a las que se refiere Bloom y, por tanto, deben formar una parte importante del constructo que ha de evaluarse. En la práctica, no es siempre fácil separar los niveles de capacidad para determinar qué nivel puede utilizar el alumno para responder a una pregunta determinada, o incluso para decir qué nivel representa la mayor exigencia educativa en un contexto dado, pero esto no libra a los sistemas de evaluación de la necesidad de abordar toda la gama. Con este propósito, el sistema de evaluación del Programa del Diploma debe incluir tareas que requieran que los alumnos reflexionen sobre sus conocimientos y realicen trabajos extensos en respuesta a la tarea establecida.

### 3.5 Gama de tareas de evaluación y de instrumentos de evaluación (componentes)

Una pregunta de opción múltiple, una pregunta de respuesta corta, una pregunta de respuesta extensa, una redacción, un proyecto, un trabajo de portafolio/carpeta, un trabajo de investigación, son todos ejemplos de tareas de evaluación. Un instrumento/componente de evaluación está compuesto de una o más tareas que se reúnen por motivos de continuidad temática o de contenido, o por conveniencia. Una prueba de examen, una carpeta, un proyecto o trabajo de investigación son ejemplos de instrumentos de evaluación o componentes. Los conceptos de tarea de evaluación y componente se solapan. A veces, un alumno puede realizar una sola tarea de entre varias a elegir, para un componente. En el sistema de evaluación del Programa del Diploma, un único examinador corrige un componente determinado correspondiente al trabajo de cada alumno.

Existen varios motivos por los que se utilizan una gran diversidad de tareas de evaluación y componentes para el Programa del Diploma. En primer lugar, desde una perspectiva histórica y pragmática, Peterson (2003) dice respecto al desarrollo inicial de la evaluación en el Programa del Diploma que “teníamos tanto una obligación como una oportunidad de tener en cuenta las diferentes técnicas de evaluación utilizadas en aquellos países a cuyas instituciones la mayoría de los alumnos del BI pretendían acceder”. En segundo lugar, la variedad de técnicas de evaluación contribuye a reducir la posibilidad de falta de equidad en la evaluación, según lo comentado en la sección 2.8 (véase también, Linn, 1992; Brown, 2002). Existen también consideraciones teóricas, relativas a la adecuación a los fines que se persiguen, que exigen un enfoque variado de la evaluación. La gama de componentes y el conjunto de tareas que incluyen garantizan que, si se realizan en todo el modelo de evaluación correspondiente a una asignatura completa (véase la sección 4.2), el logro del alumno respecto a todos los objetivos específicos de esa asignatura se encuentre adecuadamente representado.

Los objetivos específicos determinados al principio de la guía de cada asignatura definen el constructo que se evalúa correspondiente a cada una de ellas. La naturaleza de lo que debe evaluarse queda así definida con precisión para alumnos, profesores, padres/tutores y examinadores. Dado que los objetivos pueden representar muchos tipos de capacidades o destrezas distintas, las tareas de evaluación y los componentes pueden igualmente variar considerablemente en cada asignatura y de una a otra. Unos ejemplos correspondientes a dos asignaturas diferentes ilustrarán lo dicho. Primero, considérense tres de

los nueve objetivos específicos del curso de literatura, Lengua A1 de Nivel Superior (véase la sección 4.1). Son los siguientes (IBO, 1999, p. 6):

“... se espera que los estudiantes sean capaces de demostrar:

- su capacidad de realizar de manera independiente un análisis literario que revele una respuesta personal a la literatura
- la habilidad de expresar ideas con claridad, coherencia, concisión, precisión y fluidez, tanto en la comunicación oral como en la escrita
- una apreciación de las semejanzas y las diferencias entre obras literarias de diferentes períodos o culturas.”

Expresados así, éstos y los otros seis objetivos específicos ofrecen una idea muy clara tanto del tipo de destrezas que han de enseñarse como del tipo de tareas de evaluación que deben emplearse para permitir que los alumnos demuestren estas destrezas. El logro del alumno en cuanto al primer objetivo específico se demuestra a través de una exposición oral sobre un fragmento de la obra estudiada no indicado previamente y a través de una redacción de dos horas realizada en condiciones de examen sobre dos de las obras estudiadas. El segundo objetivo se aborda a través de todos los componentes de la evaluación (dos pruebas escritas, dos trabajos de literatura mundial y un comentario y exposición orales). El tercer objetivo se aborda a través de un estudio comparativo de al menos dos obras de la sección de literatura mundial del curso, realizado a lo largo de un período de tiempo y no en condiciones de examen.

La naturaleza genérica de los objetivos y las capacidades superiores que se expresan en ellos ofrecen claras indicaciones sobre el posible formato de evaluación. Aunque las tareas son abiertas, y claramente deben calificarse aplicando el criterio profesional, más que concediendo puntos de forma analítica, se ofrece orientación sustancial a profesores y alumnos sobre los parámetros de la tarea, y los examinadores deben aplicar criterios de evaluación establecidos al corregir el trabajo.

Todos los cursos de Ciencias Experimentales plantean a los estudiantes los mismos grupos de objetivos específicos, que son los siguientes (IBO, 2001b, p.7):

1. Demostrar que comprenden:
  - a) los hechos y conceptos científicos
  - b) las técnicas y los métodos científicos
  - c) la terminología científica
  - d) los métodos de presentación de la información científica.
2. Aplicar y utilizar:
  - a) los hechos y conceptos científicos
  - b) las técnicas y los métodos científicos
  - c) la terminología científica para comunicar información de forma eficaz
  - d) los métodos apropiados de presentación de la información científica.
3. Construir, analizar y evaluar:
  - a) hipótesis, problemas de investigación y predicciones
  - b) técnicas y métodos científicos
  - c) explicaciones científicas.
4. Demostrar las aptitudes personales de cooperación, perseverancia y responsabilidad que les permitirán resolver problemas y realizar investigaciones científicas de forma eficaz.
5. Demostrar las técnicas de manipulación necesarias para llevar a cabo investigaciones científicas con precisión y bajo condiciones de seguridad.

La estructura de estos objetivos sigue la taxonomía de Bloom (Bloom et ál., 1956) con bastante fidelidad. Todas las asignaturas de Ciencias Experimentales se evalúan mediante cuatro componentes: tres pruebas escritas y trabajo práctico de laboratorio, que corrige el profesor de clase. La prueba 1 consiste en un test de opción múltiple, diseñado para cubrir ampliamente el contenido del curso, evaluando los objetivos 1 y 2. La prueba 2 consiste en una pregunta de análisis de datos, algunas preguntas de respuesta corta y una pregunta de respuesta larga (dos en el Nivel Superior) basadas en el contenido del tronco común del curso que estudian todos los alumnos. La prueba 3 consiste en preguntas de respuesta corta sobre las opciones concretas que los alumnos han elegido estudiar. En las pruebas 2 y 3 las preguntas se diseñan para que su contribución a los objetivos 1 y 2, por una parte, y al objetivo 3, por otra, sean iguales. El trabajo práctico incluye la evaluación de los cinco objetivos. Por otro lado, la estructura de los exámenes garantiza que se cubra de forma equilibrada y adecuada el contenido del curso.

Los ejemplos anteriores ilustran el modo en que se establece y mantiene la validez de constructo (véase el apéndice A.1) a través de todo el Programa del Diploma. Los objetivos de aprendizaje de cada asignatura definen el constructo, proporcionan un marco para el contenido del curso, influyen significativamente en la enseñanza de la asignatura, determinan qué tareas e instrumentos de evaluación deben utilizarse, y también proporcionan a los examinadores una sólida orientación sobre qué características del trabajo del alumno deben valorarse. Los objetivos normalmente se definen en términos de destrezas o capacidades. El grado de conexión de los objetivos con el contenido prescrito del curso variará de una asignatura a otra (bajo grado de conexión directa en el caso de Lengua A1, por ejemplo, y alto grado de conexión directa en el de las asignaturas de ciencias y matemáticas), y también influirá en el formato de los componentes de evaluación y en las tareas que los constituyen.

## 3.6 Papel de la opinión profesional

Las complejas capacidades educativas superiores en las que se centra la evaluación en el Programa del Diploma no se prestan fácilmente a una corrección atomizada y deconstruida. Las respuestas de los alumnos a muchas de las tareas de evaluación pueden ser muy variadas y adoptar muchas formas igualmente correctas y válidas. A menudo no es posible ofrecer orientación precisa a los evaluadores sobre a qué corresponde exactamente cada punto que debe concederse, ni tampoco es esto deseable. La sección 2.4 se refiere a los avances en la teoría del aprendizaje, que sugieren que las capacidades y conocimientos complejos no deben enseñarse descomponiéndolos en bloques pequeños y descontextualizados. El mismo principio se aplica a las tareas diseñadas para evaluar el aprendizaje y al modo en que los examinadores deben evaluar las respuestas de los alumnos. La misma complejidad de las capacidades que se evalúan niega la posibilidad de calificarlas mecánicamente. Existen algunas áreas temáticas, por ejemplo, las matemáticas y las ciencias, en las que los enfoques de corrección analíticos constituyen la regla general, pero incluso aquí, no es posible prescribir exactamente cuál debe ser la respuesta de los alumnos, y los examinadores deben ser conscientes siempre de la necesidad de reconocer la validez de respuestas alternativas.

Mucho depende, pues, de la opinión profesional de quienes corrigen el trabajo del alumno, y especialmente de la pericia profesional de los examinadores supervisores que controlan el trabajo de todos los evaluadores (incluidos los profesores). Esto representa un importante desafío para la fiabilidad e integridad del sistema de evaluación, pero un desafío al que debemos enfrentarnos. Debe invertirse mucho esfuerzo en proporcionar formación, orientación y material de apoyo, y en desarrollar sistemas de control de calidad para comprobar los estándares de corrección (véanse las secciones 5.4 y 5.5). El personal de IBO colabora estrechamente con los examinadores supervisores en el desarrollo de estos mecanismos de apoyo y de control de calidad.

Existe esa dependencia de la opinión profesional en muchos sistemas basados en estándares, o en criterios, y en muchos sistemas basados en la competencia, que a menudo se utilizan para titulaciones de formación profesional. Gardner (1999) establece una gran diferencia entre la evaluación que se realiza mediante pruebas objetivas estandarizadas, por una parte, y la que se realiza mediante métodos de aprendizaje que tienen lugar en el contexto del trabajo práctico, por otra. La evolución del aprendizaje es juzgada por el

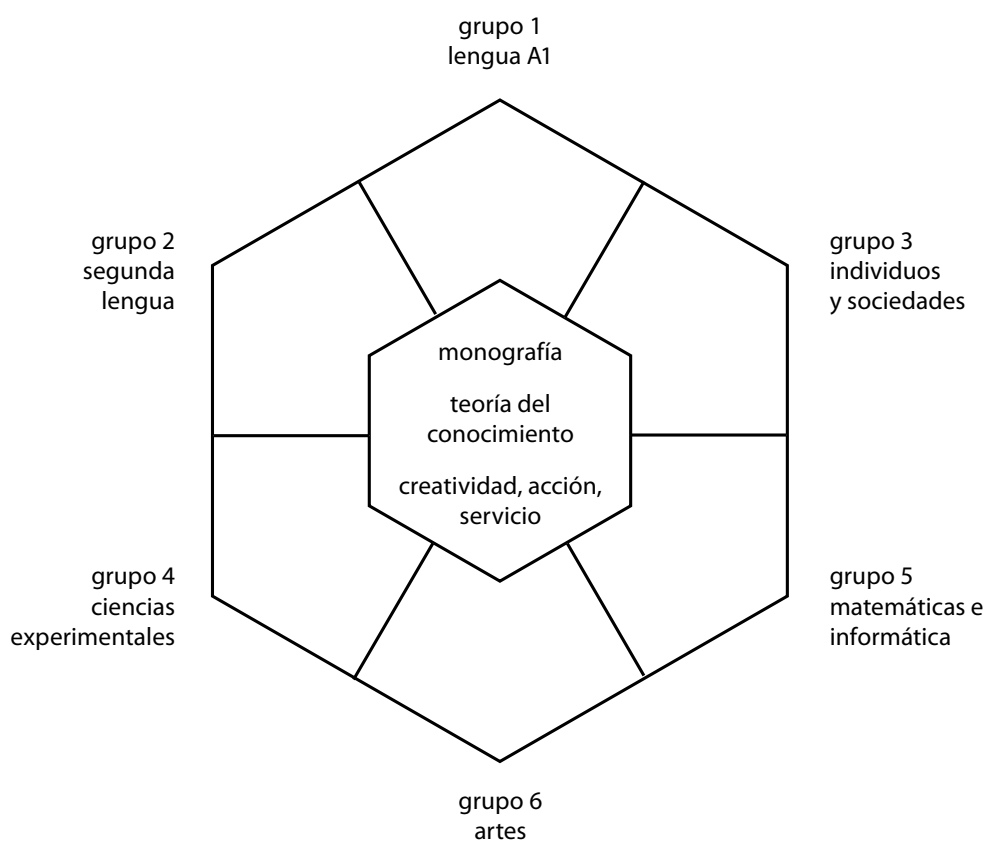
maestro, que tiene un conocimiento directo de los procedimientos de trabajo del aprendiz, así como una visión amplia de su producción en diversos contextos. Gardner rechaza con firmeza las acusaciones de que un sistema así es demasiado subjetivo, argumentando que puede demostrarse que es lo suficientemente fiable y que, en cualquier caso, las denominadas pruebas objetivas tienen en la práctica un gran sesgo a favor de los alumnos que poseen una determinada combinación de inteligencia lingüística e inteligencia lógica.

El sistema de evaluación formal adoptado por IBO para el Programa del Diploma incluye elementos de ambos extremos. Existen ciertas tareas de evaluación muy formales, entre las que se incluyen los tests de opción múltiple, pero también hay muchas más tareas abiertas sustanciales, y la evaluación se centra asimismo en el proceso, al calificar el profesor los proyectos y el trabajo práctico del alumno. Esta gama de logros del alumno se resume en una calificación final de la asignatura (en una escala de 1 a 7) según los descriptores de las calificaciones finales, que representan los estándares finales de cada asignatura. Puede que dichos estándares existan sobre el papel de forma genérica, y que se refuercen mediante materiales que incluyan ejemplos escritos, pero para el análisis final, la complejidad y diversidad de la información que ha de sintetizarse para llegar a una valoración concreta, se requiere la interpretación de los estándares que se encuentra en las mentes del equipo de examinadores supervisores con experiencia. Es cierto que los examinadores supervisores y el personal de IBO acuden a datos estadísticos para verificar sus valoraciones, pero las decisiones fundamentales a la hora de corregir y otorgar las calificaciones finales se basan en juzgar los logros del alumno con relación a los estándares previstos.

## 4 Estructuras de evaluación del Programa del Diploma

### 4.1 El currículo del Programa del Diploma

El Programa del Diploma es un curso de dos años de duración para alumnos de 16 a 19 años. Ofrece un currículo amplio, producto de la búsqueda deliberada de un equilibrio entre la especialización precoz de ciertos sistemas nacionales y la universalidad preferida en otros. El modelo curricular se presenta en forma de hexágono, con seis áreas académicas en torno a un centro. Los alumnos deben estudiar simultáneamente una serie asignaturas que representan todas las disciplinas fundamentales.



**Figura 1**

*La estructura hexagonal del currículo del Programa del Diploma.*

Los alumnos que estudian el programa completo deben elegir una asignatura de cada uno de los Grupos 1 a 5. Además, deben elegir una sexta asignatura, o del Grupo 6, o de uno de los otros grupos como segunda asignatura. Se estudian tres asignaturas (a veces cuatro) en Nivel Superior (NS) y tres asignaturas (a veces dos) en Nivel Medio (NM). Las horas lectivas recomendadas para un curso de NS son 240, y para un curso de NM, 150. Así se permite a los alumnos que profundicen más en las áreas de las asignaturas que prefieren, a la vez que se les exige que sigan estudiando en otras áreas. Los cursos de NM son versiones reducidas de cursos de NS de la misma asignatura.

Si los alumnos no son capaces de estudiar el programa completo, pueden estudiar menos cursos, y recibirán una certificación individual de los resultados de cada uno de ellos.

### **Grupo 1: Lengua A1**

El Grupo 1 está compuesto de cursos de literatura en la primera lengua del alumno. Los cursos inician a los alumnos en la literatura de distintos períodos, géneros y estilos. Los alumnos depuran sus destrezas de expresión escrita, expresión oral y análisis, y aprenden técnicas de crítica literaria. Los cursos ayudan a los alumnos a conservar fuertes lazos con su propia cultura a la vez que les ofrecen una perspectiva internacional mediante el estudio de la literatura de otras partes del mundo.

### **Grupo 2: Lengua *ab initio*, Lengua B, Lengua A2, Lenguas Clásicas**

La adquisición de una segunda lengua tiene mucha importancia en el Programa del Diploma. Los alumnos aprenden a comprender y utilizar la lengua y a profundizar en las culturas de los países donde se habla dicha lengua. Este grupo de materias incluye cursos para principiantes (Lengua *ab initio*, Lenguas Clásicas), para estudiantes de segunda lengua con experiencia previa de la lengua (Lengua B), y para alumnos bilingües con un alto nivel de fluidez (Lengua A2).

### **Grupo 3: Individuos y Sociedades**

Este grupo incluye nueve asignaturas: Economía, Geografía, Historia, Filosofía, Psicología, Antropología Social y Cultural, Empresa y Gestión, Historia Islámica, y Tecnología de la Información en una Sociedad Global. Estudiando la experiencia y conducta humanas, así como las instituciones y entornos económicos y sociales, los alumnos logran comprender distintas perspectivas y valores. Aprenden a analizar conceptos y teorías, y a utilizar métodos cualitativos y cuantitativos de recopilación y análisis de datos.

### **Grupo 4: Ciencias Experimentales**

Las ciencias que se ofrecen en este grupo son: Biología, Química, Física, Sistemas Medioambientales y Tecnología del Diseño. Los alumnos se familiarizan con el cuerpo de conocimientos, métodos y técnicas que caracterizan a las ciencias y a la tecnología, y aprenden destrezas prácticas de laboratorio.

### **Grupo 5: Matemáticas e Informática**

Este grupo incluye cursos diseñados para distintos grados de capacidad e intereses. Algunos van dirigidos a alumnos que desean profundizar en el estudio de las matemáticas, mientras que otros son para quienes necesitan las matemáticas para mejorar su comprensión de otras asignaturas. El objetivo de los cursos es proporcionar a los alumnos los conocimientos y principios de las matemáticas. Ayudan a los alumnos a desarrollar el pensamiento lógico y creativo en matemáticas y a utilizar la abstracción y la generalización para llegar a conclusiones. Todos los alumnos deben hacer un curso de matemáticas, y también pueden elegir estudiar informática.

### **Grupo 6: Artes**

El grupo de Artes incluye: Artes Visuales, Música y Artes Teatrales. Se pone el énfasis en hacer arte, es decir, los alumnos logran comprender las artes y aprenden a expresarse artísticamente creando, produciendo e interpretando obras artísticas. Además de ello, exploran formas artísticas de diferentes culturas de todo el mundo.

### **Requisitos básicos**

En el núcleo del Programa del Diploma se encuentran tres requisitos que los alumnos deben cumplir, además de los correspondientes a las seis asignaturas.

- **Teoría del Conocimiento**

Uno de los elementos más importantes del Programa del Diploma es el curso de Teoría del Conocimiento, que reta a los alumnos a cuestionarse las bases del conocimiento: a reflexionar críticamente sobre cómo conocen aquello que consideran hechos o que creen que es la verdad. Consiste casi completamente en examinar preguntas sobre las diferentes fuentes del conocimiento (la percepción, el lenguaje, las emociones, la razón) y los diferentes tipos de conocimiento (científico, artístico, matemático, histórico), tales como:

- ¿Construimos la realidad o la reconocemos?
- ¿El conocimiento siempre requiere algún tipo de base racional? ¿Existe algún tipo de conocimiento que pueda adquirirse únicamente a través de las emociones?

- ¿El conocimiento científico es progresivo/siempre ha ido en aumento? ¿Podemos llegar a un punto en el que se conozca todo lo importante en sentido científico?
- **Creatividad, Acción y Servicio (CAS)**  
Otro elemento importante del Programa del Diploma es Creatividad, Acción y Servicio (CAS). Para cumplir este requisito, los alumnos deben tomar parte en actividades artísticas (creativas); en deportes, expediciones o proyectos internacionales o locales (acción); y en proyectos de servicio comunitario o social (servicio). La participación en CAS aumenta la conciencia de los alumnos sobre las necesidades de la comunidad y les ofrece la oportunidad de aplicar lo que han aprendido en clase para abordar estas necesidades. También les da confianza en su capacidad de cambiar las cosas. Los proyectos deben tener resultados tangibles y beneficiar a otros de forma real. También es una parte importante de la implicación de los alumnos en CAS el que reflexionen sobre lo que experimentan.
- **Monografía**  
La monografía, de 4.000 palabras como máximo, ofrece a los alumnos la oportunidad de profundizar en el estudio de un tema que les merezca especial interés. La experiencia y destrezas que adquieren al investigar independientemente y producir un trabajo escrito estructurado y sustancial les proporciona una preparación excelente para el estudio independiente a nivel universitario.

La estructura curricular del Programa del Diploma define el marco en el que debe operar la evaluación. Se construyen modelos de evaluación individuales para cada asignatura tanto a NS como a NM, para Teoría del Conocimiento (TdC) y para la Monografía. Se celebran dos convocatorias de exámenes al año, en mayo y en noviembre, y los resultados se publican a principios de julio y a principios de enero respectivamente. Los resultados que se publican corresponden a las calificaciones finales de las asignaturas (otorgadas en base a una escala de 1 a 7) y a las calificaciones de TdC y la Monografía (la escala es de E, la nota más baja, a A, la más alta). Mediante una matriz de puntos (de 0 a 3) se convierte en una nota numérica las calificaciones de TdC y de la Monografía. CAS no contribuye al total de puntos, pero la certificación de la participación en CAS es un requisito sin el cual no puede concederse el diploma.

La puntuación máxima posible de un alumno del Programa del Diploma es de 45 puntos (6 × 7, más 3). Se concederá el diploma al alumno que obtenga 24 puntos o más, sujeto a ciertas condiciones relativas a la distribución de puntos en todas las asignaturas. La política de que tanto los cursos de NS como los de NM dispongan del mismo número de puntos a pesar de la diferencia de carga de trabajo y logro entre ambos niveles, es intencionada, a fin de animar a los alumnos a considerar que sus cursos de NM y sus cursos de NS son igualmente importantes. Se anima a los alumnos a alcanzar los máximos niveles en todas las disciplinas y dicho esfuerzo se recompensa adecuadamente.

## 4.2 Modelos de evaluación y papel de la evaluación interna

A fin de proporcionar lo necesario para realizar la evaluación formal del programa, existen modelos de evaluación para cada nivel (NS y NM) de cada asignatura y también para cada uno de los requisitos troncales, excepto CAS. Cada modelo comprende varios componentes de evaluación distintos, que normalmente incluyen tareas diferentes. Los modelos de evaluación son revisados como parte del proceso de revisión del currículo correspondiente a cada asignatura. Las asignaturas son revisadas cada siete años por un grupo de revisión que incluye profesores, examinadores, personal de IBO y asesores externos. El proceso de revisión del currículo es un proceso consultivo en el que se hacen circular propuestas entre los colegios autorizados para impartir el Programa del Diploma, a fin de que hagan comentarios según van desarrollándose dichas propuestas. Las recomendaciones y propuestas de los grupos de revisión del currículo también se someten a la consideración del Comité de revisión del Programa del Diploma. Dicho comité es responsable de la calidad académica global de los cursos que componen el Programa del Diploma, y aprueba los programas de las asignaturas y los modelos de evaluación propuestos. El comité tiene especialmente en cuenta:

- el estándar académico y la equiparación de los diferentes cursos
- la reducción al mínimo del solapamiento de contenidos y objetivos entre asignaturas y el fomento de cursos que se complementen unos a otros
- el control de la carga de la evaluación global que han de soportar alumnos, profesores e IBO, para garantizar que puedan sobrellevarla
- la eliminación de la duplicación innecesaria de evaluaciones.

En el contexto de la evaluación, tanto los grupos de revisión del currículo como el Comité de revisión del Programa del Diploma toman como referencia la política de evaluación del Programa del Diploma (véase el apéndice B), que define los parámetros dentro de los cuales deben desarrollarse los modelos de evaluación.

La mayoría de los componentes de la evaluación son exámenes compuestos por una gran variedad de tipos de preguntas para adaptarse a los requisitos de la asignatura. Entre ellos se encuentran: preguntas de opción múltiple (utilizadas sólo en un pequeño número de exámenes), preguntas de respuesta corta, preguntas estructuradas de resolución de problemas, preguntas abiertas de resolución de problemas, preguntas para desarrollo de un tema, preguntas de análisis de datos, estudios de casos, y comentarios sobre textos que se les dan a los alumnos. Los exámenes se realizan en un entorno controlado, y las respuestas de los alumnos son evaluadas externamente por examinadores independientes.

Existen otros componentes/tareas que los alumnos realizan guiados por sus profesores durante un período de tiempo extenso, y que también son evaluados externamente por examinadores. Por ejemplo: trabajos de literatura mundial en Lengua A1, tareas escritas en Lengua A2, investigaciones en Música, ensayos en Teoría del Conocimiento y las monografías. La evaluación de todos ellos se centra en la calidad del producto escrito acabado, lo que hace que sean adecuados para ser evaluados externamente.

Existe un tercer tipo de componente de evaluación, la evaluación interna, que consiste en que el trabajo de los alumnos es corregido por el profesor y sujeto a moderación externa (véase la sección 5.5). La política de evaluación del Programa del Diploma prevé la existencia de un componente evaluado internamente para todos los cursos en los que se considere apropiado que así sea. Son muy pocos los cursos que no tienen un componente evaluado internamente. La evaluación interna permite que se incluyan en el modelo de evaluación componentes/tareas que demuestren los logros del alumno con respecto a objetivos que no se prestan a ser examinados externamente. Esto se refiere, en particular, a destrezas de procedimiento, como las que pueden demostrarse en actividades tales como trabajos de proyecto, trabajos de campo, trabajos prácticos de laboratorio e investigaciones matemáticas. Aunque los cuadernos y las carpetas de trabajo pueden utilizarse para dejar constancia de las destrezas de procedimiento, no constituyen en sí mismos instrumentos adecuados para la evaluación externa. No obstante, proporcionan un modo de que los moderadores (examinadores externos) puedan comprobar que el estándar de la corrección de los profesores es el adecuado. La evaluación interna se utiliza también para el trabajo oral en los cursos de lengua, lo que permite a los profesores elegir el momento más idóneo para llevar a cabo la evaluación formal del trabajo oral, así como para proporcionar un entorno favorable para ello.

El trabajo evaluado internamente tiene otras ventajas en el contexto de las titulaciones internacionales. Permite mucha flexibilidad en la elección de los temas y, a la vez, seguir abordando un conjunto de destrezas comunes. Los colegios pueden así situar el estudio en un contexto cultural o geográfico local, o establecer lazos más estrechos entre el aula y el mundo exterior que la rodea. Los colegios internacionales, cuyos alumnos a menudo tienen unos antecedentes culturales distintos de aquéllos en los que se encuentra ubicado el colegio, pueden utilizar el trabajo de evaluación interna para implicarse de forma más estrecha en la sociedad o entornos locales. Por otra parte, la evaluación interna puede utilizarse de distintos modos para desarrollar vínculos con culturas lejanas, generalmente comunicándose por correo electrónico con colegios de otras partes del mundo. Brown (2002) también señala el valor de la evaluación interna del Programa del Diploma a la hora de dar cabida a la diversidad cultural, lo que fomenta “una perspectiva más amplia del internacionalismo”, permitiendo una multiplicidad de enfoques culturales y ofreciendo a cada alumno la oportunidad de experimentar diversos valores culturales.

Además, la evaluación interna a menudo ofrece a los alumnos la oportunidad de elegir sus propios temas o cuestiones, según sus intereses particulares, y les permite mayor control de su propio aprendizaje. Esta flexibilidad de enfoque hace que la evaluación interna sea una valiosa adición a la educación de los alumnos, incrementando la validez, no sólo del proceso de evaluación, sino también de la experiencia de aprendizaje en su conjunto.

La evaluación interna plantea algunas dificultades significativas como, por ejemplo, garantizar la fiabilidad y la autenticidad, y evitar la excesiva carga de trabajo. La fiabilidad de la corrección es una cuestión de especial importancia. Cuando la evaluación interna contribuye a un sistema de evaluación "de alto riesgo", puede situar al profesor de la clase en una posición difícil, a la vez como juez y parte en el aprendizaje del alumno. A este posible conflicto de intereses se le añade un fuerte elemento de subjetividad en torno a la naturaleza de la relación personal entre profesor y alumno. La opinión del profesor puede verse influida por su experiencia previa del trabajo del alumno, lo que crea ciertas expectativas. Puede que los profesores a veces no tengan claros los límites de su papel en orientar y apoyar a los alumnos mientras llevan a cabo el trabajo de evaluación interna, y con frecuencia puede que tengan sólo una visión limitada de los estándares globales de logro en el área de su propia asignatura, es decir, cuando evalúen el trabajo de sus propios alumnos pueden verse muy influenciados por los estándares generales existentes en su colegio.

Éstos son los motivos por los que los moderadores a menudo tienen que ajustar las notas de las evaluaciones internas realizadas por los profesores, incluso aunque los moderadores no tengan una visión tan completa de los logros del alumno como los profesores. Si bien IBO proporciona a los profesores materiales de apoyo sobre cómo orientar y corregir el trabajo del alumno, no está en posición de seleccionar y conservar sólo a aquellos profesores capaces de calificar de forma coherente con arreglo al estándar correcto, como puede hacer con los examinadores externos. A pesar de tales reservas, la investigación ha demostrado que la evaluación interna que llevan a cabo algunos profesores, puede ser tan fiable como la evaluación externa (Black, 1993b).

La autenticidad es otro problema que suscita preguntas sobre la fiabilidad de la evaluación interna. Debido al temor a que otras personas distintas del alumno puedan contribuir de forma significativa al trabajo que se lleva a cabo, algunos sistemas de evaluación han excluido completamente la evaluación interna, o exigido que las tareas evaluadas internamente se lleven a cabo siempre bajo supervisión en el aula. La opinión de IBO es que se trata de una exageración. Para el trabajo supervisado internamente pero calificado externamente, tanto el profesor como el alumno deben firmar una declaración de autenticidad. Los profesores deben también firmar una declaración de que el trabajo evaluado internamente es el realizado por el propio alumno. Si, posteriormente, se demuestra que el trabajo no es realmente del alumno, es posible que se considere conducta fraudulenta (IBO, 2003b). El plagio, especialmente utilizando Internet, supone evidentemente una gran preocupación, y se están tomando medidas firmes para identificar y penalizar el trabajo plagiado y disuadir a los alumnos de que plagien.

La tercera gran preocupación sobre la evaluación interna es la de la carga de trabajo que crea, especialmente para el profesor y el alumno. Internamente se establecen tareas normalmente importantes que exigen que el alumno les dedique mucho tiempo. Aunque es correcto que los profesores empleen mucho tiempo preparando a los alumnos en las destrezas y procesos necesarios para la evaluación interna, puede existir una fuerte tentación, por parte de alumno y profesor, de ensayar y practicar la tarea concreta establecida para la evaluación interna más de lo necesario, a fin de hacerlo lo mejor posible. El peligro de esto es que no quede suficiente tiempo para la enseñanza y el aprendizaje del resto del curso, al igual que puede ocurrir si se dedica demasiado tiempo a un tema concreto del contenido de la asignatura. En tales circunstancias, también puede surgir la impresión de que la evaluación interna resulta pesada y agotadora. Es importante que los alumnos aprendan cuál es el mejor modo de organizar su tiempo y planificar su propio aprendizaje. De hecho, se trata de uno de los objetivos más amplios del Programa del Diploma en su conjunto, y es parte del proceso de adoptar una actitud activa de aprendizaje.

A pesar de las reconocidas limitaciones que supone depender de la evaluación interna como medio principal de evaluar los logros del alumno, debe tenerse en cuenta que existen varios sistemas educativos nacionales y públicos que utilizan únicamente la evaluación interna sujeta a diversos grados de regulación externa

para la evaluación final del nivel secundario pre-universitario. Esto refleja el alto valor que a menudo se le concede a este modo de evaluación.

En resumen, a pesar de las muchas ventajas de incluir la evaluación interna en un sistema de evaluación formal, existen buenas razones para establecer un límite máximo a su contribución al sistema de evaluación global del Programa del Diploma, según estipula la política de evaluación. Salvo unas pocas excepciones, como son las asignaturas artísticas con su enfoque eminentemente práctico, la evaluación interna constituye una parte secundaria del modelo de evaluación correspondiente a cada asignatura.

## 4.3 Personal

Cada convocatoria de exámenes del Programa del Diploma depende totalmente del enorme esfuerzo de grandes equipos de personas. Muchos miles de profesores preparan a los alumnos y contribuyen directamente a través de la evaluación interna. Aproximadamente 4.000 examinadores se encuentran distribuidos tan ampliamente por el mundo como los profesores. Dirigiendo a los examinadores de cada asignatura hay un grupo de examinadores supervisores, un examinador jefe y uno o más examinadores jefe adjuntos, según el número de alumnos matriculados y de las necesidades de evaluación de la asignatura<sup>1</sup>. Los examinadores jefe y los examinadores jefe adjuntos normalmente ejercen el cargo durante un período de cinco años.

Es responsabilidad del grupo de examinadores supervisores preparar los exámenes y los esquemas de calificación para cada convocatoria, dirigir los equipos de examinadores, tomar parte en la reunión de evaluación y encargarse de tomar decisiones sobre los resultados como parte del servicio de consultas sobre resultados (véase la sección 5.9). Normalmente, se nombran examinadores jefe a personas que pertenecen al sector de la enseñanza superior, por los conocimientos técnicos sobre la asignatura que puedan aportar, porque conocen los requisitos para estudiar en el nivel universitario, y porque no puede haber conflicto de intereses a la hora de juzgar el trabajo del alumno. Los examinadores jefe adjuntos son, por lo general, examinadores con experiencia que también ejercen como profesores del Programa del Diploma. Aportan un alto nivel de experiencia práctica al proceso de evaluación.

A los examinadores jefe adjuntos, y a otros examinadores a quienes puede acudir ocasionalmente para ayudar en la preparación de exámenes, no se les permite trabajar en la preparación de exámenes para las convocatorias en las que sus colegios tienen a alumnos matriculados. Por lo tanto, un adjunto que es profesor en un colegio que tiene alumnos matriculados en las convocatorias de mayo sólo puede colaborar en la preparación de los exámenes de las convocatorias de noviembre. El examinador jefe, junto con otros examinadores supervisores y asesores externos no vinculados a colegios autorizados para impartir el Programa del Diploma y el personal de IBO, tienen la necesaria visión de conjunto de los exámenes de las dos convocatorias anuales como para garantizar que sean equiparables. Hay unas pocas asignaturas, sin embargo, que carecen de exámenes en el modelo de evaluación, por ejemplo, Artes Visuales, Artes Teatrales y Teoría del Conocimiento. En estos casos, el conflicto de intereses es una cuestión de menor importancia. Todos los examinadores supervisores corrigen el trabajo de alumnos de ambas convocatorias anuales, pero no se les permite corregir trabajo de ningún colegio con el que tengan conexiones personales, profesionales o geográficas.

Todos los examinadores jefe de cada asignatura forman la Junta de examinadores, que tiene un presidente electo. El presidente de la junta preside el Comité de revisión del Programa del Diploma (véase la sección 4.2) y el Comité de la evaluación final (véase la sección 5.7), y toma parte en otros comités de IBO, cabiendo señalar el Consejo de Fundación. Uno de estos comités es el Comité del Programa del Diploma, responsable

---

<sup>1</sup> Las asignaturas que tienen niveles de matriculación más bajos, generalmente las lenguas, pueden tener asignado un sólo examinador, que lo hace todo, desde la preparación de los exámenes hasta su corrección y determinación de las bandas de calificación; a este examinador se le llama "examinador responsable" de esa asignatura.

de supervisar la estructura, regulación e implementación del Programa del Diploma. Los examinadores jefe de cada grupo de materias del hexágono del Programa del Diploma eligen un representante del grupo que asiste al Comité de revisión del Programa del Diploma. Los demás miembros de este comité son personal de IBO y representantes de los colegios.

El nombramiento de examinador asistente tiene una vigencia de un año, aunque en muchos casos se les ofrece la posibilidad de prórrogas anuales durante un período de tiempo prolongado. En su mayoría, estos examinadores provienen del profesorado del Programa del Diploma de todo el mundo, pero también de otros ámbitos educativos, siempre que cuenten con la adecuada experiencia en la evaluación de la asignatura correspondiente. La actividad de corregir exámenes supone una importante oportunidad de desarrollo profesional para los profesores del Programa del Diploma, ofreciéndoles una nueva perspectiva del modo en que los profesores de otros colegios preparan a sus alumnos, así como una familiarización personal mucho mayor con los estándares de trabajo que se esperan. A fin de evitar posibles conflictos de intereses, nunca se da a corregir a los examinadores asistentes trabajo de los colegios con los que tengan alguna conexión personal o profesional.

Los examinadores supervisores trabajan muy estrechamente con los responsables de asignaturas (RdA) responsables de currículo (RdC), es decir, el personal de IBO que se encarga de dirigir la revisión del currículo y la evaluación de sus asignaturas. Cada RdA y RdC se encargará de la dirección de un pequeño número de asignaturas. El currículo y la actividad de evaluación correspondiente al Programa del Diploma se organizan desde las oficinas del Centro de currículo y evaluación (IBCA) de Cardiff, Reino Unido, que son las más grandes que tiene la organización en todo el mundo.

En Cardiff se encuentra también el departamento de producción de exámenes (DPE), en el que se reciben las versiones definitivas de los exámenes, se realiza su composición tipográfica, se les da el formato correspondiente, se supervisa su traducción, se imprimen y se envían a los colegios. También hay un departamento de administración de exámenes (DAE), que se encarga de distribuir el trabajo de los alumnos entre los examinadores y garantiza que dicho trabajo se corrija a tiempo y se devuelva a Cardiff. El personal del departamento de administración de exámenes pasa mucho tiempo trabajando con los coordinadores del Programa del Diploma (los miembros del personal de cada colegio responsables de organizar el trabajo de los exámenes) en cuestiones tales como la matriculación de los alumnos y las consultas administrativas. También mantiene contacto con examinadores sobre cómo va evolucionando el trabajo de corrección para asegurarse de que todos los datos de las puntuaciones estén disponibles para las reuniones de evaluación y que todo el procesamiento de puntos esté terminado para la publicación de los resultados. Finalmente, las cuestiones técnicas de evaluación, tales como los procedimientos de moderación y la investigación, las lleva a cabo el personal de evaluación dirigido por el director de evaluación, que tiene la responsabilidad global de todos los aspectos del sistema de evaluación formal del Programa del Diploma.

## 5 Procedimientos de evaluación del Programa del Diploma

### 5.1 Preparación de exámenes

La preparación de exámenes que tengan un alto grado de validez de constructo (es decir, que representen adecuadamente las diversas destrezas o capacidades que, según los objetivos específicos del curso, se requieren en una selección equilibrada de contenidos del mismo), que planteen un desafío intelectual del nivel adecuado, y que tengan el menor sesgo posible, supone un gran reto. Para afrontar este reto, un equipo de examinadores supervisores y personal de IBO, con la colaboración de asesores externos, emplean un largo período de tiempo en preparar exámenes nuevos para cada convocatoria. Este trabajo de preparación puede comenzar de 18 meses a un año antes de la convocatoria correspondiente a un curso determinado. A continuación se describen las principales fases del procedimiento para una asignatura típica con un nivel de matriculación alto, en el que participa un equipo de examinadores supervisores. Cuando se trata de asignaturas con niveles de matriculación bajos, un solo “examinador responsable” puede encargarse de todas las fases.

La primera fase es la de encargar los exámenes. A cada miembro del equipo de examinadores supervisores (el examinador jefe y los examinadores jefe adjuntos, y a veces otros examinadores con experiencia) se le encarga la preparación de uno o más exámenes hasta formar el lote completo que se necesita para los cursos de ambos niveles, Nivel Superior (NS) y Nivel Medio (NM), correspondientes a una asignatura. El examinador supervisor puede redactar todo el examen o puede compilar el examen a partir de las preguntas que le presenten otros examinadores. Para conservar la validez de constructo, existe una especificación para cada examen, en la que se detalla el número y tipo de preguntas que debe tener (véase la sección 3.5). Esta estructura fija contribuye también a garantizar que lo que se exige en las sucesivas versiones del examen sea equiparable. Los examinadores supervisores siempre procuran plantear preguntas que cognitivamente exijan la variedad adecuada de destrezas o capacidades, y que no sean demasiado predecibles. En cada examen se deben plantear preguntas o tareas que sean de algún modo distintas de lo que los alumnos hayan hecho antes. Aunque existe un límite para el tipo de preguntas que pueden hacerse en un curso determinado cualquiera, al menos en algunas de ellas debe pedirse a los alumnos que resuelvan problemas o que piensen de forma creativa, que apliquen lo que saben en un contexto nuevo en lugar de simplemente poner en práctica destrezas bien ensayadas o de repetir conocimientos adquiridos.

Se pone el mismo esfuerzo en la preparación del esquema de calificación que acompaña a cada examen. Incluso cuando se trata de exámenes con preguntas/tareas abiertas, que se califican según los mismos criterios de evaluación en cada convocatoria (véase la sección 5.4), se preparan notas explicativas para la corrección que sirven de orientación para que los examinadores sepan cómo aplicar los criterios en el contexto de cada pregunta. Los esquemas de calificación son tan importantes para la integridad del proceso de evaluación como los exámenes. Cada esquema de calificación es mucho más que un simple conjunto de respuestas modelo, puesto que en él se ofrece orientación sobre cómo calificar los enfoques alternativos más comunes que pueden utilizar los alumnos al responder a las preguntas, así como los errores o equivocaciones que con más frecuencia cometen.

Después de haber preparado las versiones iniciales de un lote de exámenes, el equipo examinador supervisor y el responsable de asignatura (RdA) o el responsable de currículo (RdC) correspondiente celebran una reunión de edición de exámenes en la que revisan el trabajo. En esta reunión, el lote de exámenes para cada nivel se juzga en su conjunto, para ver cómo se cubren el contenido y los objetivos específicos del curso en los diferentes exámenes que componen el lote. El grupo también hace una revisión crítica de cada pregunta considerando su adecuación, corrección y cualquier posibilidad de ambigüedad o sesgo. Se examina también minuciosamente cada pregunta y su esquema de calificación correspondiente. Los miembros del equipo examinador supervisor pueden seguir comunicándose antes y después de esta reunión en un entorno electrónico seguro.

Una vez se llega a un acuerdo sobre las versiones revisadas, las “versiones post-reunión” se envían a un asesor externo, el cual no tiene conexiones con el proceso de preparación del examen. El asesor externo ofrece una opinión independiente sobre la adecuación de cada lote de exámenes en términos de su cobertura del contenido del curso, su nivel de dificultad en comparación con lotes de exámenes anteriores, sobre si puede ejecutarse dentro del tiempo disponible, y sobre todos los factores previamente considerados en la reunión de edición de exámenes. Posteriormente, redacta un informe sobre los exámenes, que el RdA/RdC y el equipo examinador supervisor deben estudiar. El autor del examen y el RdA/RdC deben tomar conjuntamente las decisiones sobre las cuestiones suscitadas por el asesor externo, requiriéndose a veces la intervención del examinador jefe.

Cuando se llega a un acuerdo final sobre el contenido de los exámenes y esquemas de calificación, el departamento de producción de exámenes (DPE) se encarga de la planificación y administración de cada lote de exámenes, realizando la composición tipográfica, dándole el formato correspondiente, imprimiéndolo y enviándolo a los colegios. El RdA/RdC y el examinador supervisor que los redactaron, revisan cada examen, corrigen las pruebas y ofrecen comentarios en las distintas etapas. En cuanto a los exámenes con un alto nivel de contenido técnico, un revisor adicional comprueba de forma independiente las pruebas finales, respondiendo a todas las preguntas y comprobando las respuestas con referencia al esquema de calificación, para asegurarse de que no se hayan introducido o pasado por alto errores de términos técnicos, símbolos o números. Este revisor debe ser un examinador con experiencia que no haya estado implicado en el proceso de preparación del examen en absoluto hasta ese momento, para evitar que el estar familiarizado con el contenido del examen pueda obstaculizar la eficacia del proceso de corrección de la prueba.

La Tabla 1 (véase la página siguiente) muestra una relación de todas las fases de la preparación de un examen típico. Debe preverse un período de tiempo para la traducción al francés y al español de los exámenes de las asignaturas de los grupos 3 a 6. Normalmente la traducción se empieza sólo cuando las pruebas de las versiones en inglés de los exámenes han sido corregidas por el examinador y ya están casi listas, aunque las cuestiones que se susciten en el proceso de traducción, particularmente las relativas a ambigüedades imprevistas y terminología específica de la asignatura, pueden influir en la versión en inglés de los exámenes.

El calendario esquematizado en la Tabla 1 se ha elaborado teniendo en cuenta varios factores importantes. En primer lugar, para producir exámenes de gran calidad, es necesario, en muchas fases del proceso, juzgar y discutir distintas cuestiones y reflexionar concienzudamente. En segundo lugar, gran parte del personal clave involucrado (los RdA/RdC y examinadores supervisores) pueden tener otros compromisos que les impidan atender la preparación de exámenes en el momento en que los materiales estén listos para la siguiente fase. En tercer lugar, en momentos críticos del proceso, es necesario que todos los exámenes de un lote estén preparados antes de poder pasar a la siguiente fase, por lo tanto, un lote puede retrasarse por culpa de uno sólo de los exámenes. En cuarto lugar, la cantidad de exámenes que han de prepararse significa que el personal de DPE no siempre puede prestar atención inmediata a cada uno. Para la convocatoria de mayo, hay más de 700 exámenes distintos que preparar, incluidas las traducciones. Para la de noviembre, hay aproximadamente 400 exámenes. Teniendo en cuenta las dos convocatorias anuales, y la duración del calendario de preparación, es probable que el personal del departamento de preparación de exámenes llegue a estar trabajando en 2.000 exámenes en un momento dado, lo que significa que no puede prestar atención inmediata a cada uno en cuanto está listo para la fase siguiente. Igualmente, el gran número de exámenes distintos y el elevado volumen de copias que deben imprimirse de algunos de ellos, significa que los encargados de la impresión necesitan mucho tiempo para imprimir y cotejar todos los exámenes. El objetivo del calendario de producción es garantizar que los exámenes lleguen a los colegios unas tres semanas antes de que empiece el calendario de exámenes. Esto se hace en previsión de posibles retrasos en la red global de entrega, especialmente en lo relativo al despacho de aduanas. También permite a los colegios tener tiempo de comprobar las entregas y de poder rectificar posibles errores.

No se realizan pruebas previas de los exámenes debido a los recursos que se necesitarían para ello y a la dificultad que supondría encontrar grupos experimentales adecuados.

Fase de producción	Duración (en semanas)
1. Encargo de exámenes y preparación de versiones iniciales	14
2. Entrega de las versiones pre-reunión a DPE: formateo de documentos	5
3. Reunión de edición de exámenes, y revisión	4
4. Entrega de las versiones post-reunión a DPE: modificación y comprobación	4
5. Entrega de versiones post-reunión a consejero externo: realización de informe	5
6. Comentarios adicionales de RdA/RdC	3
7. Entrega de versiones finales de los examinadores a DPE/RdA/RdC	9
8. Comprobación de versiones finales por RdA/RdC	3
9. Modificaciones finales, formateo y comprobación por RdA/RdC	4
10. Corrección de pruebas y comprobación adicional por DPE	3
11. Corrección final de pruebas por examinador	4
12. Modificaciones finales y correcciones finales por RdA/RdC	4
13. Revisión por revisor adicional y posibles modificaciones	7
14. Comprobación final de estilo (de la organización) y de coherencia interna	1
15. RdA/RdC dan por terminado el proceso; envío a imprenta	3
16. Devolución de pruebas de imprenta para comprobación	1
17. Impresión de exámenes y embalaje para envío a colegios	12
<b>Tiempo total</b>	<b>86 semanas</b>

DPE: Departamento de producción de exámenes

RdA: Responsable de asignatura

RdC: Responsable de currículo

**Tabla 1**

*Calendario típico de producción para un examen del Grupo 4 o del Grupo 5.*

## 5.2 Los exámenes

Los exámenes de cada convocatoria tienen lugar durante un período de unas tres semanas en mayo y en noviembre. Dado el número de asignaturas objeto de examen, deben realizarse a veces exámenes de dos o tres asignaturas distintas en el mismo horario para ceñirse al calendario. Las asignaturas de estos bloques se eligen de modo que se reduzca al mínimo la probabilidad de que haya alumnos a los que les puedan coincidir asignaturas. No obstante, es inevitable que así ocurra en algunos casos, y el procedimiento para resolverlos se describe en el *Vademécum* (manual de procedimiento para coordinadores y profesores del Programa del Diploma).

Los exámenes se programan para evitar, siempre que sea posible, que haya más de seis horas de exámenes en un día en circunstancias normales. Los exámenes tampoco tienen lugar los viernes por la tarde a fin de mostrar cierta consideración hacia aquellos colegios cuya semana lectiva no es de lunes a viernes. La norma general que se aplica a los exámenes de un curso concreto es que las dos o tres pruebas de que se compone la evaluación externa de una asignatura se tomen seguidas, empezando a la tarde y acabando a la mañana siguiente. Esta distribución es preferible a la de organizar todos los exámenes de un curso determinado el mismo día, por los siguientes motivos:

- Para la mayoría de los alumnos, esto supone una distribución más uniforme de los exámenes durante el calendario de tres semanas, y les permite recuperarse en caso de que crean que no han dado todo lo que pueden de sí mismos en un momento dado.
- Si un alumno está enfermo o no puede evitar ausentarse un día determinado, todavía le queda la oportunidad de hacer los exámenes pendientes el día antes o el día después. Si se han cubierto suficientes componentes de evaluación, puede que aún sea posible concederle una calificación en ciertas circunstancias.
- Los exámenes, una vez respondidos por los alumnos, se envían a los examinadores en días distintos, reduciéndose así la posibilidad de que todos ellos puedan extraviarse en el trayecto.

No todos los exámenes pueden seguir la pauta descrita anteriormente si se ha de completar el programa en tres semanas sin utilizar los viernes por la tarde. Los exámenes de Lengua A1 y Lengua A2 se organizan de modo que entre la prueba 1 y la prueba 2 haya un período de varios días. Dado que estas pruebas son bastantes independientes entre sí en cuanto a su contenido, y que se trata de una lengua en la que el alumno tiene mucha fluidez, se considera que la separación de las pruebas tendrá un efecto menor que en otras asignaturas.

Los colegios deben realizar los exámenes siguiendo unas reglas muy estrictas estipuladas en el *Vademécum*. Estas reglas lo cubren todo, desde gestionar la recepción de los materiales de examen, pasando por la celebración de los exámenes, hasta enviar los exámenes a los examinadores para su corrección. El personal de la oficina regional de IBO y los asesores llevan a cabo visitas de inspección aleatorias a los colegios durante el calendario de exámenes para comprobar los procedimientos administrativos y las medidas de seguridad del colegio.

El *Vademécum* y una publicación que aborda los problemas de los alumnos con necesidades especiales de evaluación ofrecen información a los colegios sobre cómo organizar este tipo de evaluación, por ejemplo, para alumnos con problemas de aprendizaje, de conducta, físicos, sensoriales, médicos, o de salud mental.

### 5.3 Evaluación interna y otros componentes no de examen

La evaluación interna puede adoptar diversas formas, desde una exposición oral individual y una conversación de diez minutos de duración para los cursos de Lengua B, hasta una carpeta de investigación en Artes Visuales, que constituye un registro personal del desarrollo artístico del alumno, a la que se recomienda que se dediquen 72 horas de trabajo en el Nivel Superior (casi un tercio del curso). Entre estos extremos se encuentran casos como la evaluación interna de las Ciencias Experimentales (Grupo 4), compuesta en el Nivel Superior de trabajos elegidos de una carpeta de 60 horas de prácticas e investigaciones (25% del total de horas lectivas). La naturaleza de la tarea de evaluación refleja el objetivo de la evaluación interna, en particular su énfasis en las destrezas de procedimiento y en el tipo de las mismas. Éste es el caso, en especial, del Grupo 4, en el que deben elegirse del conjunto de trabajos prácticos aquéllos que cumplan ciertos criterios.

Sin embargo, todas las evaluaciones internas presentan ciertas características de procedimiento comunes. En primer lugar, la evaluación interna debe, en la medida de lo posible, integrarse en la enseñanza normal de clase. La evaluación interna se centra en las destrezas o capacidades, no en el contenido de la asignatura,

pero las actividades de evaluación interna, elegidas ya sea por el profesor o por el alumno, a menudo pueden utilizarse como vehículos para la enseñanza del contenido prescrito del curso. Las actividades que se utilizan para la evaluación interna también pueden servir para desarrollar destrezas o capacidades, es decir, formativamente, así como para contribuir sumativamente al resultado final de la evaluación. La decisión sobre cuándo pasar de utilizar una actividad para la evaluación formativa a utilizarla como parte de la evaluación sumativa final, a menudo corresponde al profesor. La evaluación interna no debe considerarse como una actividad separada “de cierre” que ha de llevarse a cabo después de que se haya impartido el curso.

La segunda característica común se refiere al nivel de apoyo que el profesor debe prestar al alumno en actividades que contribuyen a la evaluación final. Cuando el resultado final de la actividad es un trabajo escrito relativamente formal, generalmente se permite a los profesores discutir el tema y el enfoque con el alumno y aconsejarle, con ciertas restricciones, sobre la primera versión. Todas las modificaciones o correcciones posteriores debe realizarlas el alumno, de modo que el trabajo final que se presente para la evaluación interna sea el realizado por él mismo. A veces se permite que el trabajo evaluado internamente pueda basarse en actividades de grupo, pero siempre que deba presentarse trabajo escrito, ha de ser el realizado individualmente por cada alumno.

En tercer lugar, la evaluación interna se lleva a cabo aplicando un conjunto de criterios de evaluación determinado correspondiente a cada curso (véase la sección 5.4). Estos criterios describen los tipos y niveles de destrezas o capacidades que son objeto de la evaluación interna. Los profesores deben asegurarse de que los alumnos estén familiarizados con dichos criterios, y de que los trabajos seleccionados para la evaluación interna efectivamente los aborden. Esto es particularmente importante en el Grupo 4, Ciencias Experimentales, en el que la carpeta de prácticas puede incluir varios trabajos que, aunque perfectamente admisibles, no sean adecuados con respecto a los criterios de evaluación. Los criterios de evaluación interna del Grupo 4 deben aplicarse a un conjunto de destrezas o capacidades determinado que puede no resultar evidente en ciertos trabajos estándar de laboratorio de ciencias.

Las últimas dos características se aplican también al pequeño número de componentes no de examen corregidos externamente, que incluyen las monografías, los ensayos de Teoría del Conocimiento, los trabajos de literatura mundial de Lengua A1, las tareas escritas de Lengua A2 y las investigaciones de Música. Aunque estos trabajos se envían a los examinadores y no los corrigen los profesores, el papel de éstos en discutir el trabajo con el alumno, en aconsejarle y en tener en cuenta los criterios de evaluación es muy similar al de los componentes de evaluación interna que sí corrigen los profesores.

## 5.4 El trabajo de corrección

Se ha mencionado anteriormente la importancia de la fiabilidad del evaluador, es decir, que un proceso de evaluación conceda casi los mismos puntos a un trabajo determinado con independencia de quien lo evalúe y del momento en el que se evalúe (véase el apéndice A.3). Principalmente, hay tres formas de garantizar esto. En primer lugar, es importante nombrar y conservar sólo a aquellos examinadores que corrijan de forma sistemática y objetiva. La gran mayoría de examinadores del Programa del Diploma son profesores del programa con experiencia. Estos examinadores son perfectamente idóneos para desempeñar la tarea, dado que ya están familiarizados con el curso que se está impartiendo, con los requisitos de evaluación correspondientes, y tienen un cierto conocimiento de los estándares previstos. En segundo lugar, en cada convocatoria se comprueban las puntuaciones concedidas por todos los examinadores, excepto las del examinador supervisor de cada componente: el que hayan hecho un buen trabajo en convocatorias anteriores no garantiza que vuelvan a hacerlo en la siguiente. Este procedimiento se llama moderación, y se describe en la sección 5.5. El tercer método, que aquí se examina más detalladamente, consiste en instruir a los examinadores de manera exhaustiva sobre cómo proceder en cuanto a la corrección. Puede hacerse mediante formación previa, un área de actividad que IBO tiene planeado desarrollar mucho más en un futuro próximo por medios electrónicos. Se instruye detalladamente a los examinadores del Programa

del Diploma sobre los procedimientos administrativos que han de seguirse para hacer posible que la moderación se realice con éxito, y también reciben información importante sobre cómo asignar los puntos.

Para esto último IBO emplea principalmente dos métodos: esquemas de calificación analíticos y criterios de evaluación (que tienen una variante llamada bandas generales de calificación).

### 5.4.1 Esquemas de calificación analíticos

Los esquemas de calificación analíticos se preparan para aquellas preguntas de examen que se espera que los alumnos contesten con un tipo concreto de respuesta y/o una respuesta final determinada. Estos esquemas de calificación instruyen específicamente a los examinadores sobre cómo desglosar la puntuación total disponible para una pregunta con respecto a las diferentes partes de ésta, reflejando la importancia que el equipo examinador supervisor da a esas partes distintas de la pregunta. Los alumnos pueden acertar o equivocarse en distintas partes de la pregunta, y las preguntas extensas estructuradas están diseñadas de modo que si un alumno se equivoca al principio de la pregunta, el resto de la misma no le resulte inaccesible. Ésta es la razón principal para utilizar preguntas estructuradas en asignaturas técnicas como las ciencias y las matemáticas: permitir que los examinadores puedan conceder puntos por aciertos parciales en las respuestas. Si una pregunta con profundidad carece de estructuración, es posible que algunos alumnos no puedan conseguir muchos puntos debido a algún pequeño error al principio de su respuesta, o por haber entendido algo ligeramente mal y haber seguido una dirección equivocada. Los esquemas de calificación analíticos más elaborados se encuentran en las asignaturas de Matemáticas. Estos esquemas contienen instrucciones específicas sobre cómo corregir tipos concretos de respuestas incorrectas, y cómo seguir hasta el final el proceso iniciado por el alumno cuando ha cometido un error en una parte de la pregunta.

Incluso con preguntas estructuradas en las que se esperan respuestas muy específicas, los esquemas de calificación deben ofrecer a los examinadores suficiente información para que puedan calificar de forma coherente los principales tipos de enfoques distintos que puedan adoptar los alumnos, y los errores más corrientes que puedan cometer. Los exámenes contendrán siempre algunas preguntas en las que los examinadores deberán utilizar su propio criterio profesional a la hora de distribuir los puntos en respuestas inesperadas o que constituyan alternativas válidas, pero los esquemas de calificación deben proporcionar toda la orientación posible sobre cómo ejercer dicho criterio.

Además de esquemas de calificación y criterios de evaluación, los examinadores asistentes reciben asesoramiento de los examinadores supervisores por teléfono o correo electrónico durante el período de corrección propiamente dicho.

Puede que cuando el equipo examinador supervisor redacte un examen y su esquema de calificación correspondiente no siempre consiga prever todos los posibles tipos de respuestas más corrientes de los alumnos. Aunque se hace todo lo posible, es especialmente difícil predecir la gama de respuestas a nivel global, dada la diversidad de culturas educativas y estilos de enseñanza que existen en el mundo. Para abordar este problema y reducir al máximo la dependencia de la opinión, quizás variable, del examinador, los examinadores supervisores de cada examen correspondiente a asignaturas con un alto nivel de matriculación se reúnen poco después de que se haya celebrado el examen para revisar los exámenes de una muestra de alumnos. Esto se denomina reunión de estandarización, y su objetivo es realizar adiciones o modificar a última hora el esquema de calificación a la vista de las respuestas reales de cierto número de alumnos, y asegurarse de que los examinadores supervisores estén de acuerdo en cuanto a la aplicación del esquema de evaluación. Esto es de crucial importancia para el éxito de la moderación, y las cuestiones que surgen de esta reunión se comunican directamente a todos los examinadores asistentes.

## 5.4.2 Criterios de evaluación

Cuando una tarea de evaluación es tan abierta que la diversidad de respuestas válidas posibles es demasiado amplia como para poder preparar esquemas de calificación analíticos que las cubran todas, se aplican entonces criterios de evaluación. Los criterios de evaluación no se refieren al contenido específico de la respuesta, aunque algunos puedan referirse a la necesidad de que los alumnos demuestren conocimientos concretos del contenido. Se concentran más en las destrezas o capacidades genéricas que los alumnos deben demostrar, con independencia de los aspectos individuales específicos de la respuesta. Por ejemplo, los cinco criterios de evaluación para el examen de Lengua A1 que requieren un comentario escrito de uno o dos pasajes de un texto no estudiado previamente se titulan: comprensión del texto, interpretación del texto, apreciación de los rasgos literarios, presentación y uso del lenguaje. Cada criterio comprende un conjunto de destrezas o capacidades relacionadas, y los alumnos deben demostrar su nivel de logro en cada una de ellas. Los criterios para evaluar la tarea escrita de Lengua B (el curso principal de lengua extranjera) son: corrección y fluidez lingüísticas, interacción cultural (es decir, estilo, registro, recursos, y estructura apropiados al destinatario al que vayan dirigidos) y comunicación del mensaje. Como tercer ejemplo, los cuatro criterios de las preguntas de ensayo de Filosofía basadas en los temas opcionales de estudio son: claridad de expresión, conocimiento y comprensión de las cuestiones filosóficas, identificación y análisis de material pertinente, y desarrollo y evaluación.

Debido a su naturaleza muy variable, las evaluaciones internas y las tareas no de examen evaluadas externamente también se califican utilizando criterios de evaluación. Para la evaluación interna de todas las Ciencias Experimentales del Grupo 4, hay ocho criterios:

- planificación inicial de la investigación (definición del problema, formulación de hipótesis y selección de variables)
- selección de equipo y diseño de un método
- obtención de datos
- procesamiento y presentación de datos
- conclusión y evaluación
- técnicas de manipulación (uso de técnicas y seguimiento de instrucciones)
- aptitudes para el trabajo en equipo
- motivación y manera ética de trabajar.

Existe una estrecha relación entre estos criterios y los objetivos del curso que se exponen en la sección 3.5, lo que apoya el alto grado de validez de constructo (véase el apéndice A.1).

En todos los casos en que se aplican criterios de evaluación existen descriptores de nivel de logro para cada criterio, que definen las diferencias en los niveles de logro susceptibles de merecer puntuaciones distintas y describen las formas típicas de medir las respuestas de los alumnos con relación a cada criterio. Se llega a la puntuación total que es posible conceder a un trabajo determinado sumando el nivel de logro máximo correspondiente a cada criterio. Se les da mayor peso a los criterios que se consideran más importantes, atribuyéndoles un mayor número de niveles de logro.

Es importante tener en cuenta que, aunque los descriptores de nivel de un criterio tienen una naturaleza jerárquica, y a menudo se refieren a la jerarquía de las capacidades cognitivas definidas por Bloom et ál. (1956), ambas jerarquías son independientes entre sí. Los descriptores de nivel más bajos no se refieren sólo a las capacidades cognitivas "más simples", ni los descriptores del nivel más alto se reservan únicamente para las capacidades cognitivas "de orden superior". Se reconoce que existen diferentes niveles de logro en cada una de las áreas de las capacidades cognitivas.

Los siguientes niveles de logro correspondientes al criterio "interpretación del texto" para la prueba del comentario escrito de Lengua A1 (IBO, 1999, p. 44) sirven de ejemplo. Este criterio mide la pertinencia de

las ideas del alumno sobre el texto, si ha explorado esas ideas de manera satisfactoria, si ha ilustrado sus afirmaciones de forma satisfactoria, y hasta qué punto ha expresado una respuesta personal pertinente. Los niveles de logro ilustran la estrechísima conexión existente entre el primer objetivo determinado para Lengua A1 en la sección 3.5 y el proceso de evaluación que se lleva a cabo al final del curso. El objetivo es el siguiente: "...se espera que los estudiantes sean capaces de demostrar su capacidad de realizar de manera independiente un análisis literario que revele una respuesta personal a la literatura."

### Nivel de logro

**0 El estudiante no ha alcanzado el nivel 1.**

**1 Escasa interpretación del texto**

- las ideas del estudiante no resultan, en su mayor parte, significativas o pertinentes, **o bien**
- el comentario consiste fundamentalmente en la narración o repetición del contenido.

**2 Cierta interpretación del texto**

- las ideas del estudiante son a veces irrelevantes
- el comentario consiste, principalmente, en generalizaciones no fundamentadas, **o bien**
- el comentario es, en su mayor parte, una paráfrasis del texto.

**3 Interpretación satisfactoria del texto**

- las ideas del estudiante son generalmente pertinentes
- el análisis es apropiado y está ilustrado adecuadamente mediante algunos ejemplos pertinentes.

**4 Buena interpretación del texto**

- las ideas del estudiante son claramente pertinentes e incluyen una respuesta personal apropiada
- el análisis es, en líneas generales, detallado, y ha sido bien ilustrado con ejemplos pertinentes.

**5 Excelente interpretación del texto**

- las ideas del estudiante son convincentes e incluyen una respuesta personal apropiada y madura
- el análisis es siempre detallado y está sustentado, de forma convincente, con ejemplos cuidadosamente escogidos.

El enfoque de la evaluación en el Programa del Diploma para aplicar los niveles de logro de un criterio es el modelo del "que mejor se ajuste". El examinador o profesor que aplica un criterio de evaluación debe elegir el nivel de logro que mejor coincida en su conjunto con el trabajo que está puntuando. No es necesario que se cumpla cada aspecto detallado de un nivel de logro para conceder dicho nivel, y es conveniente tener en cuenta que el nivel más alto de cualquiera de los criterios no representa la perfección como probablemente lo haría la puntuación máxima de un esquema de calificación analítico (los sistemas de calificación analíticos operan en una gama de puntuación mucho más amplia que los criterios de evaluación).

Hay varias tareas de evaluación que se califican siguiendo los mismos criterios de evaluación en todas las convocatorias. Aunque la naturaleza general de la tarea (normalmente una redacción o trabajo escrito extenso) sea la misma en todas las convocatorias, los requisitos específicos de cada pregunta pueden influir en el modo en que deben aplicarse los criterios de evaluación. Los esquemas de calificación analíticos no son adecuados en estos casos, pero el examinador supervisor que preparó el examen generalmente redacta notas para la corrección del mismo. Estas notas ofrecen a los examinadores asistentes orientación sobre como deben aplicarse los criterios de evaluación a cada pregunta. Cuando se utilizan los criterios de evaluación

en la evaluación interna, tanto los profesores como los moderadores deben tomar como referencia los materiales publicados de apoyo al profesor, que ofrecen ejemplos de aplicación de los criterios.

### 5.4.3 Bandas generales de calificación

A veces no se considera adecuado separar los diferentes criterios de evaluación aplicables a un trabajo concreto. El modo más adecuado de aplicar los criterios de evaluación es con relativa independencia unos de otros, sin que el rendimiento del alumno en un criterio se vea influido por los otros, aunque en la práctica rara vez puede conseguirse totalmente. Si surge la situación en la que no es posible discernir criterios de evaluación separables, entonces debe adoptarse un enfoque distinto. Esto también puede ser necesario cuando el trabajo que ha de evaluarse sea tan variable que del mismo no pueda derivarse un conjunto de criterios, cada uno de los cuales sea directamente aplicable a todas las respuestas. En tales casos se utilizan bandas generales de calificación en lugar de distintos criterios por separado. Estas bandas representan, de hecho, un único criterio holístico que se aplica al trabajo juzgado en su conjunto. Dado que se necesita una gama de puntuación que sirva para diferenciar razonablemente los niveles de rendimiento del alumno, a cada descriptor de nivel de la banda de calificación le corresponderán un número de puntos.

Los descriptores tienden a ser bastante extensos en sí mismos para que cubran diversas cualidades que pueda presentar el trabajo del alumno y, una vez más, deben relacionarse directamente con los objetivos del curso. Pueden encontrarse ejemplos de bandas de calificación en la guía de Historia del Programa del Diploma (IBO, 2001a). Como sucede con los criterios de evaluación, se utiliza el enfoque del “que mejor se ajuste”, debiendo los examinadores aplicar, además, su criterio sobre qué puntuación concreta conceder de entre las posibilidades que les ofrece la gama correspondiente a cada descriptor de nivel, dependiendo de cómo se ajuste el trabajo del alumno a dicho descriptor. Por ejemplo, un nivel de banda de calificación puede cubrir la gama de 6 a 10 puntos. El examinador concederá al trabajo del alumno la puntuación de esa gama que mejor se ajuste al descriptor de nivel pertinente. La investigación ha demostrado que, cuando se han aplicado métodos de evaluación basados en bandas de calificación holísticas y en criterios de evaluación a trabajos de redacción que se prestan a ambos métodos de puntuación, existe poca diferencia entre ambos, en términos de fiabilidad de la corrección (Wood, 1991, cap. 5).

### 5.4.4 Calendario de corrección

Los examinadores llevan a cabo la corrección dentro de un plazo de tiempo muy ajustado. Hay seis semanas entre la fecha del último examen y la fecha de publicación de los resultados. En el período que sigue a la fecha de celebración del examen ha de hacerse lo siguiente: debe completarse el esquema de calificación, si es necesario; los exámenes deben llegar a manos de los examinadores, que pueden perfectamente encontrarse en diferentes partes del mundo, para su corrección; estos exámenes deben ser corregidos y devueltos a IBCA; debe enviarse una muestra de los exámenes puntuados por cada examinador a un examinador con más experiencia, que los modera y los envía a IBCA; debe celebrarse la reunión de evaluación; y deben volver a corregirse aquellos exámenes que se considere necesario. La moderación de las notas de cada examinador se basa en una muestra de los primeros exámenes que corrige. La moderación de la muestra se lleva a cabo mientras el examinador continúa corrigiendo el resto de exámenes que tiene asignados. El proceso de moderación en sí mismo se describe en la sección siguiente, pero debe tenerse en cuenta que, tras haber enviado su muestra para moderación, existen métodos adicionales para comprobar la coherencia de las notas concedidas por los examinadores en todos los exámenes que tienen asignados.

En primer lugar, en la reunión de evaluación, el equipo de examinadores supervisores estudia una amplia selección de exámenes para determinar los límites de las bandas de calificación. Aunque en este proceso se pone el énfasis en valorar el trabajo del alumno más que en la puntuación, saldrán a relucir los casos de puntuaciones anómalas. En segundo lugar, después de establecer los citados límites, los examinadores supervisores vuelven a corregir el trabajo de aquellos alumnos que por muy poco no han alcanzado una calificación más alta. En esta fase, pueden detectarse más casos de puntuaciones incorrectas. En tercer lugar, en el proceso de volver a corregir el trabajo del alumno, denominado “consultas sobre los resultados” (un servicio que pueden solicitar los colegios después de que se hayan expedido los resultados), pueden una

vez más surgir indicios de que el examinador haya corregido de manera incorrecta o asistemática. Existen, por tanto, otras formas de comprobar la calidad de la corrección realizada por el examinador, además de la muestra para la moderación.

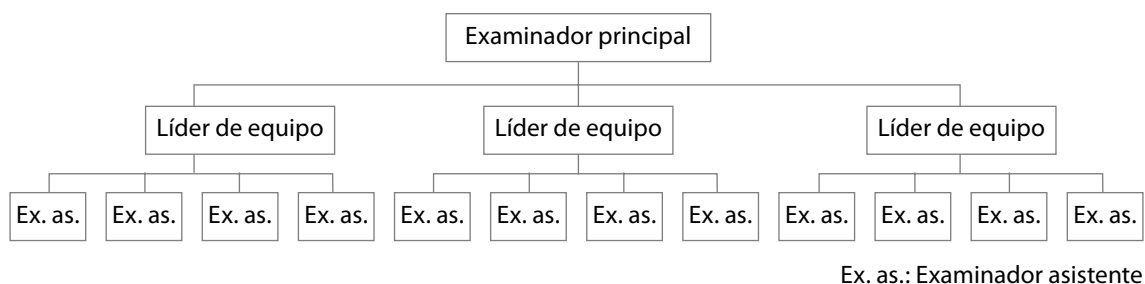
### 5.4.5 Examinadores visitantes

Finalmente, se ha hecho referencia a lo largo de esta sección a la corrección a distancia de materiales de evaluación que llevan a cabo los examinadores. Generalmente se trata de exámenes a los que los alumnos responden por escrito, pero también pueden ser cintas de video y de audio. En ciertas circunstancias, sin embargo, los examinadores visitan el colegio y evalúan el trabajo del alumno directamente. Debido a los elevados costos y a las dificultades de mantener la fiabilidad entre examinadores al intervenir un número reducido de ellos y estar muy diseminados por todo el mundo, sólo se utilizan examinadores visitantes cuando son clara e inequívocamente necesarios. Actualmente, sólo se utilizan para Artes Visuales y para un curso piloto de Danza. La necesidad de ver el trabajo de los alumnos *in situ* y comentarlo directamente con ellos supera cualquier otra consideración. Los examinadores visitantes basan sus decisiones en los criterios de evaluación, y la moderación de estas decisiones se lleva a cabo por medio de fotografías o videos. Esto supone llegar a un compromiso entre las necesidades de una evaluación fiable y las de una evaluación válida de los logros alcanzados por el alumno. Los procedimientos de corrección por medio de examinadores visitantes están sujetos a una revisión constante.

## 5.5 Moderación

### 5.5.1 Procedimiento de moderación

La moderación es el principal instrumento para garantizar la fiabilidad de la corrección, aunque no el único (véase la sección 5.4.4). Todo examinador, excepto el examinador principal de cada componente, que establece el estándar, envía una muestra del trabajo que ha corregido a un examinador con mucha experiencia, que hace la función de líder de equipo. Éste vuelve a corregir la muestra y la comparación estadística de ambas puntuaciones es la que determina si es aceptable la puntuación del examinador, quizás con algún ligero ajuste, o si es inaceptable. La moderación es un proceso jerárquico, y en la Figura 2 se ilustran los diferentes niveles de la jerarquía correspondientes a un componente típico de examen con un nivel de matriculación alto.



**Figura 2**

*Jerarquía de moderación correspondiente a un componente típico de un examen con un nivel de matriculación alto evaluado externamente. Se ha simplificado el diagrama para facilitar su presentación: la mayoría de los equipos tienen en realidad unos 10 examinadores asistentes.*

Con frecuencia, el examinador principal de un componente es el examinador jefe o el examinador jefe adjunto, pero también puede serlo un ex líder de equipo con mucha experiencia. Generalmente, el examinador principal es también el autor del examen o ha estado muy involucrado en la preparación

del mismo. Los líderes de equipo son examinadores con experiencia que han demostrado a lo largo de varias convocatorias que saben corregir de forma coherente y precisa. Los líderes de equipo generalmente supervisan un grupo de hasta 10 examinadores asistentes. El tamaño de la muestra para la moderación correspondiente a los componentes de los exámenes es el 15% del total de exámenes que tiene asignados cada examinador, es decir, entre un mínimo de 10 y un máximo de 20. Se pide a los examinadores que las muestras que presenten cubran varios colegios y una gama de puntuaciones lo más completa posible.

### 5.5.2 Criterio de correlación

Los pares de puntuaciones correspondientes a cada examen de la muestra se someten a análisis estadístico. Una de las medidas estadísticas es el coeficiente de correlación (se utiliza el coeficiente de correlación de momento-producto). Mide la coherencia de la relación entre las puntuaciones concedidas por los dos examinadores. Un coeficiente de correlación de cero indica que no existe relación en absoluto; un coeficiente de uno indica que existen una coherencia perfecta en la relación entre las puntuaciones concedidas por los dos examinadores y que coinciden en la clasificación de los alumnos de mejor a peor (aunque no es necesario que los dos examinadores hayan concedido exactamente las mismas puntuaciones). Un coeficiente de  $-1$  indica que las opiniones de los dos examinadores son sistemáticamente opuestas respecto a los méritos relativos del trabajo de los alumnos, elaborando los examinadores clasificaciones contrarias.

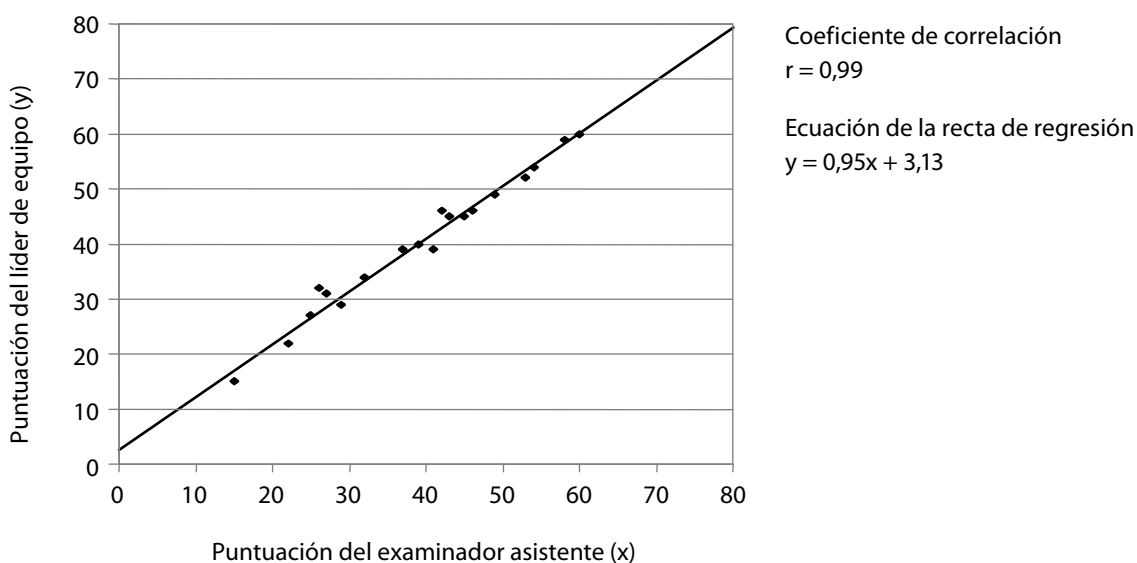
Para que sea aceptable la puntuación de un examinador, el coeficiente de correlación debe ser de al menos 0,90, lo que indica un alto nivel de coincidencia entre el examinador asistente y el líder de equipo. Si el coeficiente de correlación es inferior a 0,90, lo más probable es que los exámenes asignados al examinador asistente vuelvan a ser corregidos por otro examinador más fiable. En dichas circunstancias, no se utilizarán de nuevo los servicios de este examinador asistente a no ser que pueda determinarse una causa concreta que hubiese dado lugar a la falta de coincidencia y que ésta pudiese ser fácilmente subsanada.

No obstante, un alto coeficiente de correlación no es suficiente por sí mismo para que la puntuación del examinador sea aceptable. Es posible que un examinador asistente conceda, por ejemplo, cuatro puntos más de lo correcto a todos los alumnos en un examen puntuado sobre 25, lo que generaría una correlación de la clasificación muy alta, pero no indicaría que se ha corregido correctamente.

### 5.5.3 Regresión lineal

Se lleva a cabo otro análisis de los datos de cada muestra para moderación, lo que permite realizar un ajuste medio de todas las puntuaciones concedidas por el examinador, basado en la tendencia general que representa la muestra. La técnica utilizada se llama regresión lineal, y supone calcular la recta que mejor se ajuste a la colección de puntos de los datos correspondientes a las puntuaciones que el examinador asistente y el líder de equipo han concedido a las muestras. La Figura 3 ilustra lo expuesto anteriormente.

La ecuación de la recta de regresión obtenida a partir de los datos de la muestra puede utilizarse para convertir cada puntuación (x) concedida por el examinador asistente en una puntuación (y) equivalente, que es la que con mayor probabilidad habría concedido el líder de equipo a ese mismo alumno. Este ajuste de moderación, basado en extrapolar una muestra de puntuaciones a una colección mucho más amplia de éstas, debe necesariamente derivarse de las tendencias generales que evidencia la corrección. No pueden representarse variaciones individuales relativas a alumnos concretos. Por ejemplo, uno de los puntos de la Figura 3 muestra que el líder de equipo concedió una puntuación de 46 a un examen al que el examinador asistente concedió 42. No obstante, tal ajuste es distinto de la tendencia general, y probablemente no sería adecuado para todos los alumnos a los que el examinador asistente hubiese concedido 42 puntos. En lugar de ello, la recta de regresión, que refleja la tendencia media de la diferencia de puntuación, moderaría a 43 cada puntuación de 42 dada por el examinador asistente. Por este motivo, se llevan a cabo más comprobaciones de las puntuaciones, especialmente en “casos límite” (véase la sección 5.6). El objetivo de la moderación es garantizar que, en su conjunto, las puntuaciones de los alumnos se ajusten a niveles más apropiados. La moderación no puede garantizar el resultado exactamente correcto para cada uno de los alumnos.



**Figura 3**

*Recta de regresión de la moderación para un examinador asistente, correspondiente a un examen calificado sobre 80. Cada punto representa el par de puntuaciones que han concedido a un examen de muestra el examinador asistente y el líder de equipo. La recta de regresión continua se utiliza para convertir las puntuaciones del examinador asistente en puntuaciones moderadas.*

### 5.5.4 Otros criterios

Además de lograr un coeficiente de correlación satisfactorio, las puntuaciones de cada examinador deben cumplir otros dos criterios antes de que pueda aplicarse automáticamente un ajuste de moderación. La pendiente de la recta de regresión debe estar entre 0,5 y 1,5. Si la recta tiene muy poca pendiente significa que el examinador asistente ha espaciado las puntuaciones de los alumnos demasiado, concediendo comparativamente muy pocos puntos al trabajo flojo y demasiados puntos al bueno, aunque puede haberlo hecho de manera coherente. El líder de equipo ha tenido que comprimir considerablemente la gama de puntuaciones del examinador asistente. Se considera que la recta tiene demasiada pendiente si ésta es mayor de 1,5 y entonces significa lo contrario: el examinador asistente no ha diferenciado lo suficiente el trabajo flojo del bueno, y el líder de equipo ha tenido que expandir la gama de puntuaciones concedida.

El segundo criterio establece que la diferencia entre la puntuación media de la muestra del examinador asistente y la del líder de equipo debe ser menor del 10% de la puntuación total disponible para ese componente. Por tanto, si la puntuación total disponible para un componente es de 30, la puntuación media del examinador asistente debe estar dentro de un margen de 3 puntos de diferencia con la puntuación media del líder de equipo con respecto a una muestra determinada de trabajo. El examinador cuya puntuación deje de cumplir cualquiera de estos dos criterios puede ser coherente, pero está claramente en discordancia con los estándares previstos y se considera, por tanto, un caso de moderación fallida.

### 5.5.5 Moderación fallida

Todos los casos de moderación fallida son revisados uno por uno por el personal de evaluación de IBCA, el cual considera detenidamente los datos subyacentes y puede decidir:

- 1 aplicar algún otro tipo de ajuste de moderación, quizás ajustes lineales distintos para distintas partes de la gama de puntuaciones
- 2 requerir más datos de muestra a fin de clarificar la tendencia
- 3 requerir que se corrija de nuevo, completa o parcialmente, el trabajo asignado al examinador.

En todos estos casos, el RdA/RdC formulará la recomendación de seguir utilizando o no los servicios del examinador en futuras convocatorias.

### 5.5.6 Moderación compuesta

Debe tenerse en cuenta que los líderes de equipo también deben presentar una muestra de su trabajo de corrección y que a sus puntuaciones podrá aplicarse un ajuste de moderación mediante regresión lineal. Las ecuaciones de moderación derivadas de los distintos niveles de la jerarquía ilustrada en la Figura 2 se combinan para proporcionar un ajuste de moderación general que se aplica a cada examinador asistente. Esta combinación funciona satisfactoriamente siempre que exista un alto grado de coincidencia en los niveles superiores de la jerarquía de moderación, es decir, que el estándar de corrección de los líderes de equipo sea muy similar al del examinador principal.

### 5.5.7 Ajuste del modelo lineal

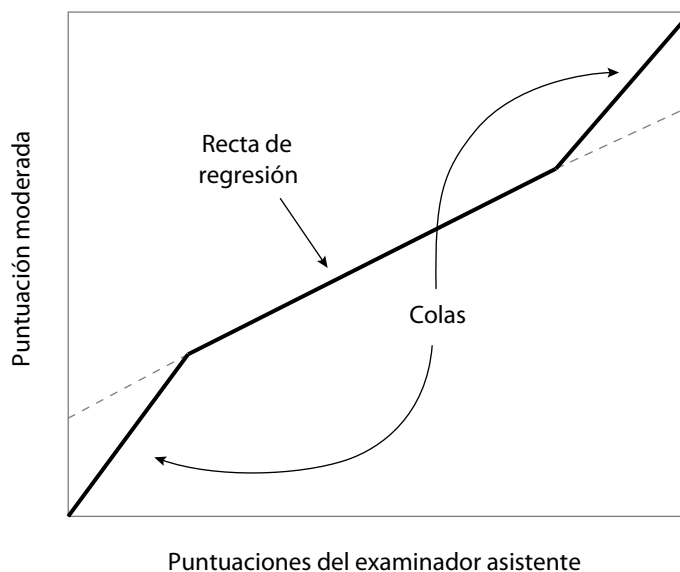
Puede cuestionarse que, tal y como se supone en el sistema actual de ajuste de moderación, la recta sea el mejor modelo para realizar el ajuste. IBO está actualmente considerando la posibilidad de utilizar una curva de ajuste óptimo en vez de una recta de ajuste óptimo. Antes de poder realizar este cambio, se deben llevar a cabo pruebas exhaustivas para asegurarse de que no se deriven funciones curvilíneas inadecuadas y que las funciones curvilíneas puedan combinarse satisfactoriamente en todos los niveles de moderación. Entre tanto, se modifica hasta cierto punto el modelo de recta utilizando “colas”. Puede observarse que un ajuste de moderación de línea recta puede producir efectos inapropiados en los extremos de la gama de puntuaciones posibles, impidiendo que pueda concederse a ningún alumno la puntuación máxima o la de cero (véase, por ejemplo, la Figura 3). A través de la moderación, puede llegar a calificarse con un cero el trabajo de un alumno que, aunque flojo, merezca unos pocos puntos o, como sucedería en el caso de la Figura 3, pueden concederse unos pocos puntos a un alumno, cuando realmente merece un cero por no haber escrito nada en su examen que merezca ningún punto.

En el extremo opuesto sucede lo mismo: un alumno cuyo trabajo sea muy bueno pero contenga algunas deficiencias evidentes podría obtener la puntuación máxima, o no obtenerla el alumno que verdaderamente la merezca. Generalmente, el impacto que tiene el ajuste de moderación basado en una recta es mayor en los extremos de la gama de puntuaciones, mientras que dicho ajuste ofrece la máxima confianza en la parte media de esa gama, donde tienden a concentrarse la mayoría de los datos de la muestra.

Para superar este problema, se aplican “colas” con relación a las puntuaciones correspondientes al 20% superior y al 20% inferior de la gama de puntuaciones disponibles. En estos extremos, se modifica la recta de regresión calculada y se sustituye por nuevas rectas o “colas” que unen la recta de regresión a las coordenadas máximas y mínimas, según se muestra en la Figura 4.

La razón fundamental para implementar este procedimiento es que, con independencia de lo generoso o estricto que pueda ser un examinador asistente, no puede calificar por debajo del mínimo ni por encima del máximo establecidos. Un examinador generoso, por ejemplo, podría acabar concediendo la puntuación máxima a alumnos que no se lo merecen del todo, así como a quienes sí se lo merecen. El proceso de moderación debe tratar por igual a todos los alumnos a los que corresponda una misma puntuación y es mejor conceder a estos alumnos el beneficio de la duda cuando se ha comprimido un segmento de logro en una única puntuación.

De este modo, los alumnos a los que el examinador asistente ha concedido la puntuación máxima (que son pocos e incluso pueden merecer una puntuación mayor que la máxima) conservan dicha puntuación máxima, cualquiera que sea el ajuste de moderación, y ningún otro alumno que corresponda a dicho examinador puede obtener esa puntuación máxima. Para ello se supone que las puntuaciones concedidas por el examinador asistente son lo bastante buenas para pasar el proceso de moderación automática. Si no lo es, no se aplican “colas”. En el extremo opuesto, una puntuación moderada de cero sólo puede derivarse de una puntuación inicial de cero. La “cola” evita que se califique con un cero el trabajo que merece unos pocos puntos, y que obtenga unos pocos puntos el que no merece ninguno.

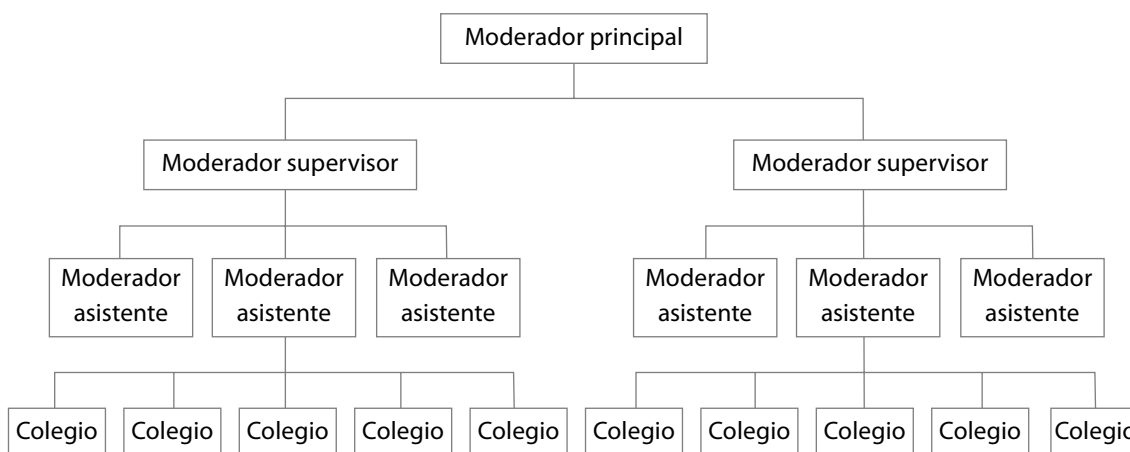


**Figura 4**

“Colas” de una recta de regresión para evitar que el ajuste desvíe la puntuación de un alumno aproximándola a los valores mínimos o máximos o distanciándola de ellos.

### 5.5.8 Moderación de la evaluación interna

La estructura básica que se sigue para que las puntuaciones iniciales, que en la evaluación interna realizan los propios profesores de clase, puedan superar o no la moderación es similar, aunque con algunas diferencias en cuanto a los criterios. (Téngase en cuenta que el superar la moderación no significa que la puntuación inicial haya sido totalmente exacta; sólo significa que ha sido lo bastante coherente como para que se le puedan aplicar los ajustes lineales de la moderación.) La Figura 5 muestra la jerarquía de moderación correspondiente a la evaluación interna.



**Figura 5**

Jerarquía de moderación que se aplica a los componentes evaluados internamente. Se ha simplificado el diagrama. En realidad, lo más probable es que cada moderador asistente reciba muestras de evaluación interna de unos diez colegios, y cada moderador supervisor tenga que supervisar a unos 10 moderadores asistentes.

La distinta denominación que reciben los examinadores que llevan a cabo la moderación de la evaluación interna (es decir, moderador principal, moderador supervisor y moderador asistente) indica el enfoque ligeramente distinto del papel de ésta. La gran mayoría de moderadores son profesores del Programa del Diploma con experiencia. Todos los componentes evaluados internamente se califican aplicando criterios de evaluación, y en la mayoría de los casos el profesor tiene acceso a muchísima más información sobre el contexto y el proceso subyacentes al trabajo del alumno que pueda tener el moderador. Por ello, se pide a los moderadores de la mayoría de los componentes de la evaluación interna, excepto en los exámenes orales de lengua, que juzguen si la puntuación del profesor parece apropiada, en lugar de simplemente volver a corregir el trabajo sin tener en cuenta la puntuación concedida por el profesor. Las puntuaciones de los profesores sólo deben modificarse cuando el moderador esté seguro de que no son adecuadas.

A pesar de este enfoque más flexible, la gran mayoría de las moderaciones fallidas corresponden a la evaluación interna. IBO no acredita a profesores con respecto a su competencia para llevar a cabo la evaluación interna, y no está en posición de controlar qué profesores la realizan, como puede hacerlo con los examinadores asistentes. También puede resultarle difícil al profesor presentar una muestra de moderación satisfactoria que cubra una amplia gama de puntuaciones o tipos de trabajo, dada la naturaleza limitada y, a veces, no representativa y muy homogénea de los grupos de alumnos de una clase de un colegio determinado. Si la muestra se encuentra inevitablemente limitada, resulta más difícil que dicha muestra cumpla los requisitos estadísticos que permitan procesarla a través del sistema de moderación automático.

Debido a estos factores, y a que el trabajo evaluado internamente es inherentemente más abierto y más susceptible de que exista subjetividad al calificarlo, los criterios para que supere la moderación automática son algo menos rigurosos. El coeficiente de correlación debe ser al menos 0,85. La pendiente de la recta de regresión debe situarse entre 0,5 y 1,5, igual que para las evaluaciones externas, pero no hay límite para la diferencia en las medias de la muestra. Se conservan las “colas” de la recta de regresión en el extremo inferior de la gama de puntuación, pero no se aplican en el extremo superior, porque es más frecuente que los profesores hayan concedido puntuaciones máximas a alumnos que claramente no las merecen. Los tamaños de las muestras de moderación son más reducidos para los componentes evaluados internamente (los colegios han enviado muestras de diez, ocho, cinco, o incluso menos trabajos a los moderadores asistentes, dependiendo del número de alumnos en cada grupo del colegio.) Esto se debe, en parte, a que muestras más grandes darían como resultando el volver a evaluar la mayor parte de los trabajos de los alumnos y, en parte, al menor número de trabajos que tienen asignados los moderadores y los moderadores supervisores. El trabajo para evaluar internamente correspondiente a un solo alumno es generalmente más voluminoso y su revisión exige más que la de un examen, lo que significa que deben asignarse menos trabajos a cada moderador. IBO selecciona a los alumnos cuyos trabajos constituirán las muestras de moderación de la evaluación interna después de que el colegio haya presentado formalmente las puntuaciones correspondientes a su evaluación interna.

Cuando hay suficientes matriculados en un curso determinado en un colegio como para que puedan repartirse en más de una clase, y más de un profesor realiza la evaluación interna, IBO espera que los profesores compartan la evaluación interna y colaboren para estandarizar el modo de aplicar los criterios. Se solicita del colegio una única muestra de moderación, que con toda probabilidad contendrá trabajo de alumnos corregido por los diferentes profesores implicados. No obstante, cuando haya en un colegio distintas clases que utilicen diferentes lenguas para responder a la misma asignatura, se requiere una muestra de moderación distinta para cada lengua.

## 5.6 Totalización y concesión de calificaciones finales

### 5.6.1 Procedimiento

La reunión de evaluación representa la culminación del proceso de evaluación de cada asignatura. Tiene lugar unos 35 días después de la fecha de los exámenes, cuando deben haber concluido los procesos de corrección y moderación de la asignatura. Las reuniones de evaluación requieren que se reúna en IBCA el equipo examinador supervisor de cada asignatura que tenga un nivel razonablemente alto de matriculación. El equipo, junto con el RdA/RdC, revisa cómo han funcionado los componentes de evaluación correspondientes a la convocatoria, establece las bandas de calificaciones finales de cada uno de los cursos de Nivel Superior y de Nivel Medio, resuelve cualquier cuestión que haya pendiente sobre las puntuaciones de los examinadores, y realiza las comprobaciones de puntuaciones de lo que constituyen “casos límite” (véase la sección 5.6.5).

La primera tarea de la reunión de evaluación es reflexionar sobre cómo ha funcionado cada componente. Todos los participantes habrán estado involucrados activamente en la corrección de al menos un componente, y la mayoría habrán contribuido a la redacción de los exámenes. Además de su propia experiencia, los examinadores supervisores comentan las observaciones sobre los exámenes presentadas formalmente por los profesores y los informes de los examinadores asistentes sobre la naturaleza de las respuestas de los alumnos en general. El equipo supervisor se sirve de esta importante información sobre la impresión general que se tiene de los exámenes para tomar sus decisiones sobre las bandas de calificaciones finales adecuadas, y también la utiliza para redactar el informe de la asignatura, que se distribuye a todos los colegios (véase la sección 5.9).

A continuación, el equipo estudia cada componente para el que deben establecerse nuevas bandas de calificaciones finales en cada convocatoria: ha de tenerse en cuenta que las bandas correspondientes a los componentes evaluados internamente y a los componentes no de examen puntuados externamente no se establecen en cada convocatoria sino, normalmente, sólo una vez, cuando se acaban de introducir o de revisar. En cada convocatoria se establecen nuevas bandas para cada prueba de examen, si bien generalmente sólo se realizan cambios menores en los límites de dichas bandas, dado que se hace todo lo posible para que cada nueva versión de una prueba tenga un nivel de dificultad global prácticamente igual a la anterior.

### 5.6.2 Establecimiento de bandas de calificaciones finales

Puede argumentarse que en un sistema basado en criterios que depende de la opinión profesional de los examinadores, los examinadores supervisores deberían poder establecer las bandas de calificación simplemente teniendo en cuenta las preguntas del examen y las respuestas que se esperan de los alumnos. Sin embargo, en la realidad es muy difícil que dichas opiniones puedan formularse con precisión antes de conocer con suficiente detalle cómo han respondido realmente los alumnos. A veces, una pregunta que se planteó creyendo que sería bastante difícil, les resulta fácil de responder a los alumnos, o viceversa. La realización de pruebas previas de las preguntas de examen puede contribuir a reducir este factor de impredecibilidad, pero para formular opiniones sobre el auténtico nivel de dificultad de una pregunta, aún debería normalmente esperarse a disponer de los resultados reales de las puntuaciones.

El establecimiento de bandas de calificaciones finales es un asunto de gran envergadura que exige muchas deliberaciones y consensuar información proveniente de diferentes fuentes: opiniones de examinadores supervisores con experiencia, comparaciones estadísticas, y expectativas de profesores con experiencia, familiarizados con los estándares y quienes mejor conocen a los alumnos.

Las bandas de calificaciones finales más importantes para cada examen, las correspondientes a las calificaciones finales 3 y 4, 6 y 7, y 2 y 3, se determinan en ese orden, siguiendo el consenso de opiniones. Se trata de las que más influyen en el progreso del alumno hacia las enseñanzas superiores. El resto se determina por interpolación a partir de estas bandas establecidas siguiendo el consenso de opiniones.

La forma principal de establecer bandas de calificación siguiendo el consenso de opiniones es revisando la calidad del trabajo del alumno con relación a los descriptores de las calificaciones finales, que son descripciones genéricas del estándar de trabajo que se espera del alumno típico respecto a cada calificación final. Los descriptores intentan definir el rendimiento típico que se espera de los alumnos con relación a una calificación final determinada, y se pretende que sirvan de orientación a los profesores de clase sobre cómo preparar a sus alumnos y cómo predecir cuáles puedan ser sus resultados finales. Sin embargo, ha de reconocerse que dichos descriptores sólo pueden interpretarse plenamente a la luz de una amplia experiencia en cuanto a su aplicación a trabajos de evaluación reales. A continuación figura un ejemplo de descriptor para una calificación final de 4 en un curso de Matemáticas:

### **Calificación 4 Satisfactorio**

El alumno demuestra un conocimiento satisfactorio del programa de estudios. Aplica principios matemáticos al realizar algunas tareas de rutina. Lleva a cabo procesos matemáticos en contextos que no implican dificultad. Muestra cierta capacidad para reconocer patrones y estructuras. Utiliza técnicas de resolución de problemas en situaciones de rutina. Tiene una comprensión limitada de la significación de los resultados e intenta extraer algunas conclusiones. Comunica las matemáticas de forma adecuada, usando algunas técnicas, notación y terminología apropiadas. Utiliza la tecnología satisfactoriamente.

Antes de la reunión de evaluación, el equipo examinador supervisor, y en especial los examinadores principales, presentarán bandas de calificación provisionales para cada componente, indicando las puntuaciones que deben incluir las bandas, basándose en su experiencia anterior de los estándares previstos. Estas bandas provisionales, junto con el consenso de la reunión sobre cómo ha funcionado cada examen y un breve estudio de la distribución de las puntuaciones moderadas (histograma) correspondientes al componente en comparación con las de años anteriores, llevan al equipo a decidir las puntuaciones que considera que debe abarcar cada banda de calificación. Se estudian varios exámenes de cada puntuación sucesivamente hasta que la reunión decide que, por ejemplo, 27 puntos representan la puntuación más alta para la calificación final de 3, y 28 puntos la más baja para la calificación final de 4.

Durante este proceso, se pide que los examinadores supervisores se centren, no en las puntuaciones del examen, sino en la naturaleza de las respuestas reales de los alumnos y hasta qué punto se corresponden con los descriptores de las notas finales. Para esta revisión, se eligen aquellos exámenes cuyas respuestas tengan, en la medida de lo posible, un nivel uniforme, en lugar de aquéllos que hayan obtenido puntuaciones altas en algunas preguntas y bajas en otras. Una vez que la reunión ha llegado a un acuerdo sobre la banda, se repite el proceso para el resto de bandas y demás componentes. Los miembros de la reunión tienen acceso a exámenes de anteriores convocatorias que representan exactamente los límites de las bandas de calificación, para ayudarles a que los estándares de las convocatorias posteriores sigan siendo coherentes.

La única excepción son las pruebas de opción múltiple. La experiencia demuestra que es muy difícil juzgar en pruebas compuestas solamente de preguntas de opción múltiple cuáles puedan ser las bandas de calificación basadas en la calidad del trabajo del alumno. Puede que esto sea así porque las respuestas contienen muy pocos datos sobre lo que los alumnos realmente han hecho y, por tanto, se carece de suficiente base para tomar una decisión. Las bandas de calificación correspondientes a dichas pruebas se calculan de modo que incluyan el porcentaje de alumnos de cada calificación final que más se aproxime al establecido siguiendo el consenso de opiniones en la prueba con la que guarden una relación más estrecha. Por ejemplo, en las Ciencias Experimentales del Grupo 4, la prueba 1 es una prueba de preguntas de opción múltiple. La prueba 2 contiene diferentes tipos de preguntas, pero se basa en el mismo contenido del curso que la prueba 1. Por tanto, si las bandas determinadas para la prueba 2 dan como resultado, por ejemplo, 9% de alumnos con una calificación final de 7, entonces se calculan las bandas correspondientes a la prueba 1 de modo que también tengan una cifra lo más cercana posible al 9% de alumnos con una calificación final de 7. De este modo, el rendimiento del alumno en la prueba de opción múltiple influye poco en la distribución global de calificaciones finales de la asignatura, pero sí influirá claramente en los resultados que cada alumno obtenga en dicha asignatura.

### 5.6.3 Totalización

Cuando se han establecido bandas de calificaciones finales para todos los componentes, se calculan las bandas provisionales de calificaciones finales de las asignaturas y se examina la distribución provisional de calificaciones finales de la asignatura. Para totalizar las puntuaciones (y bandas) de los diferentes componentes y llegar a una puntuación total porcentual, puede que primero deban ajustarse proporcionalmente. El ajuste proporcional se lleva a cabo para preservar el valor de ponderación previsto para cada componente, en cuanto a su contribución a la evaluación total del curso. Por ejemplo, un curso de Nivel Superior de una asignatura puede tener tres componentes, y el modelo requerir que el componente 1 contribuya en un 50% al resultado final, el componente 2, en un 30% y el componente 3, en un 20%. Si se diseña el componente 2 para que tenga un total de 90 puntos disponibles, entonces, tras la moderación, deben ajustarse proporcionalmente estos puntos dividiendo entre tres para lograr la ponderación del 30% requerida. Lo mismo se aplica a las bandas de calificación establecidas para el componente, que también deben calcularse partiendo de 90 puntos.

Un factor significativo aquí es que el concepto de ponderación se refiere a los resultados posibles del alumno (representados por los puntos disponibles) y no a la distribución o reparto de los puntos concedidos. La ponderación de un componente del 30% no significa que este componente deba contribuir en un 30% al reparto final de los puntos del alumno; este enfoque es más propio de los sistemas normativos, que tienen como objetivo primordial de cada ítem de evaluación y de cada test el diferenciar a los alumnos. En el sistema de evaluación del Programa del Diploma, la diferenciación entre alumnos es una cuestión de segundo orden si se la compara con la de los logros de cada alumno. El componente que no separa a los alumnos en cuanto a la puntuación que obtienen pero que, sin embargo, refleja los logros educativos significativos, contribuye valiosamente a juzgar el rendimiento del alumno. Los distintos componentes del modelo de evaluación del Programa del Diploma correspondientes a una asignatura y a un nivel determinado pueden hacer que las puntuaciones de los alumnos se repartan de diferentes maneras. Es posible que distintos componentes tengan distribuciones de calificaciones finales (el porcentaje de alumnos correspondiente a cada calificación final) bastante distintas.

Tampoco es necesario que un componente con una determinada ponderación deba contribuir exactamente en la medida de dicha ponderación a las puntuaciones totales de los alumnos. La importancia de la ponderación reside en que indica la proporción del valor global que corresponde a un determinado componente, no necesariamente la proporción global del logro.

Puede que el enfoque que adopta el sistema de evaluación del Programa del Diploma no refleje los métodos más sofisticados de ponderación, combinación (totalización) y ajuste proporcional que describe, por ejemplo, Wood (1991, cap.10), pero se basa en el sólido principio de que no pueda modificarse la puntuación de ningún alumno únicamente sobre la base de cómo le haya ido al resto de compañeros. Este enfoque también tiene la importante ventaja de la transparencia si lo comparamos con modelos más complejos de ponderación y ajuste proporcional.

Tras aplicar todos los ajustes proporcionales necesarios, las puntuaciones y los límites de las bandas de calificación se redondean al número entero más próximo. Si los decimales son exactamente 5 décimas, las cifras de las puntuaciones se redondean al alza y las de los límites de las bandas de calificación a la baja, para favorecer a los alumnos. A continuación, se suman las puntuaciones más altas correspondientes a cada calificación final de cada componente, para obtener la puntuación más alta que corresponderá a esa calificación en la asignatura. Se elige la puntuación más alta correspondiente a la calificación final con preferencia a la más baja para compensar el denominado efecto "de regresión a la media". En términos sencillos, el efecto de regresión a la media establece que es más difícil conseguir una determinada nota en varios componentes que en uno sólo. Por ejemplo, sería posible tener un modelo de evaluación con tres componentes, en el que en cada componente un 10% de los alumnos obtuvieran una calificación final de 7, mientras que menos del 5% de los alumnos obtuvieran una calificación global de 7, dependiendo de la coherencia del rendimiento del alumno en los tres componentes.

La Tabla 2 ofrece un ejemplo de totalización. Según los datos de la tabla, los límites para las calificaciones finales 6/7 en la asignatura serían las puntuaciones totales de 82/83. Por tanto, al alumno que obtuviese la puntuación más alta para la calificación final 6 en dos componentes, y la puntuación más baja para la calificación 7 en el tercero, se le concedería una calificación final de 7 en la asignatura, y así sucesivamente. Vale la pena insistir en que la calificación final que obtiene un alumno en la asignatura se determina a partir de la agregación de las puntuaciones de los componentes, y no de las notas finales de los componentes. Dado que cada nota de componente representa una gama de puntuaciones, es bastante posible que dos alumnos con las mismas notas en los componentes obtengan diferentes calificaciones finales en la asignatura.

Banda de calificación	Componente 1 (50%)	Componente 2 (30%)	Componente 3 (20%)	Límites de puntuación para la calificación final de la asignatura
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.

**Tabla 2**

*Ejemplo de agregación de bandas de calificación ajustadas proporcionalmente en el Programa del Diploma del BI.*

En todas las asignaturas del Programa del Diploma se sigue el principio de compensación de los componentes. Esto significa que si un alumno obtiene una puntuación baja en un componente, puede compensarlo con una puntuación más alta en otro. No hay "obstáculos" que superar en determinados componentes para lograr ciertas calificaciones finales en la asignatura, aparte del requisito de que el alumno debe presentar trabajo para evaluar en todos los componentes a fin de que le pueda calificar la asignatura. Por tanto, puede ser teóricamente posible, aunque muy improbable, que un alumno obtenga cero puntos en un componente y, a pesar de ello, consiga una calificación final de 7 en la asignatura, si las puntuaciones de los otros componentes son lo bastante altas. (Téngase en cuenta, sin embargo, que este mismo principio de compensación no se aplica cuando se combinan los resultados de las asignaturas para decidir si se le concede o no el diploma al alumno. En este caso se aplica un sistema de "obstáculos" por el que se exigen unas notas mínimas en TdC y en la Monografía.)

#### 5.6.4 Distribución de calificaciones finales

Tras realizar la totalización de las puntuaciones de los componentes y de las bandas de calificaciones finales mediante un proceso informático, la reunión de evaluación revisará la distribución provisional de calificaciones finales de la asignatura, antes de confirmar las bandas definitivas. Deben llevarse a cabo comparaciones con los resultados de años anteriores, con las distribuciones de notas finales previstas por los colegios y con las expectativas generales del equipo examinador supervisor. Si se produce una modificación significativa de la distribución de calificaciones finales de la asignatura en comparación con la del año anterior, deberá explicarse el motivo. Puede que haya habido un incremento importante de alumnos matriculados, produciéndose así un cambio en el nivel general de logro. En este caso puede

resultar útil analizar el rendimiento relativo de los nuevos colegios. También puede aportar datos útiles comparar el rendimiento en aquellos componentes cuyas bandas de calificación hayan permanecido fijas año tras año, como ocurre en la evaluación interna. Si la reunión de evaluación lo considera necesario, es posible volver atrás y revisar cualquier banda de calificación de cualquier componente para el que se hayan decidido ya las puntuaciones límite.

Una vez los participantes en la reunión de evaluación, incluyendo el personal de IBO correspondiente, están satisfechos de que la distribución general de calificaciones finales refleja con justicia los logros de los alumnos con referencia a los descriptores de calificaciones finales y a los resultados de años anteriores, se calculan los resultados de cada colegio y se imprimen. Se lleva a cabo una última comprobación de la adecuación de los resultados, comparando las calificaciones que han obtenido varios colegios con mucha experiencia (en los que se sabe que los profesores están muy familiarizados con los estándares requeridos) y las calificaciones que tenían previstas. Una vez más, pueden llevarse a cabo revisiones si existen discrepancias importantes.

### 5.6.5 Comprobación de “casos límite”

Cuando los resultados finales se consideran, en general, justos y correctos, el equipo examinador supervisor y otros examinadores con experiencia resuelven las cuestiones pendientes relativas a la fiabilidad de la corrección. Puede que todavía haya unos pocos examinadores cuyo trabajo deba volver a calificarse, o que acabe de descubrirse que pueda no ser fiable aunque haya sido satisfactoria la muestra de moderación. No obstante, el proceso de volver a corregir se centrará principalmente en el área de alumnos que constituyen “casos límite”. Generalmente, se trata de alumnos cuya calificación final está dos o más puntos por debajo de lo previsto, y se encuentran a dos puntos porcentuales de obtener una mejor calificación final en la asignatura. Dado que existe un margen de error en la corrección y en la moderación, debe confirmarse la exactitud de la puntuación en estos casos límite, y se volverán a corregir todos los componentes evaluados externamente de estos alumnos. Sin embargo, si resulta evidente que un colegio ha sido excesivamente optimista en muchas de las calificaciones finales que tenía previstas en la asignatura y nivel correspondientes, es probable que se tenga menos interés en volver a corregir el trabajo de dicho colegio.

Una segunda categoría de “casos límite” mucho más reducida en número, es la de los alumnos que se encuentran a sólo un punto por debajo de la calificación final prevista, y a dos puntos porcentuales de obtener dicha calificación final, y que tienen en su examen al menos un componente corregido externamente por un examinador cuyas puntuaciones no fueron plenamente satisfactorias pero cuyos exámenes no se volvieron a corregir.

Lo ideal sería que se revisasen los exámenes de todos los alumnos que se encontrasen a dos puntos de los límites de las bandas de calificación de la asignatura, para comprobar si se les ha concedido la calificación final correcta. Sin embargo, la limitación de recursos no permite que pueda hacerse esto, y por ello se centra la atención en aquellas categorías de alumnos que con mayor probabilidad puedan haber salido perjudicados en el proceso de puntuación y moderación.

### 5.6.6 Personal de apoyo

Las reuniones de evaluación cuentan con la colaboración de gran cantidad del personal de IBO, incluidos los RdA/RdC, el personal de evaluación y el personal de administración de exámenes. También hay muchos ayudantes temporales que reciben, ordenan, comprueban y trasladan todos los exámenes respondidos por los alumnos que se reciben en IBCA. Una de las obligaciones más importantes del personal temporal es comprobar cada examen para asegurarse de que todos los examinadores han calificado todas las respuestas, han concedido puntuaciones dentro de los límites asignados para cada pregunta/tarea, y han sumado, transcrito e inscrito la puntuación total correctamente.

Para contribuir a que quienes no participan directamente en las reuniones de evaluación tengan más claros los procedimientos que se siguen, se invita a profesores observadores a que asistan a ellas a fin de que informen a sus colegas sobre dicha experiencia.

## 5.7 Comité de la evaluación final

El Comité de la evaluación final se reúne después de que se hayan celebrado todas las reuniones de evaluación y justo antes de que se publiquen los resultados a principios de enero/julio. Este comité concede formalmente los diplomas y certificados a aquellos alumnos que han cumplido los requisitos. También autoriza las acciones oportunas con respecto a:

- los casos de alumnos con discapacidades, o que hayan sufrido un accidente o una enfermedad, y que no se les pueda aplicar el procedimiento normal
- los casos de alumnos que se hayan visto afectados por circunstancias imprevistas
- los casos de presunta conducta fraudulenta
- las recomendaciones del Comité de necesidades educativas especiales sobre la organización de la evaluación para alumnos con necesidades educativas especiales.

El presidente de la Junta de examinadores preside el Comité de la evaluación final, que está compuesto por un número reducido de otros examinadores jefe y personal supervisor de IBO. Se invita a un observador de un colegio a que asista a las reuniones de evaluación.

El comité se basa en la política establecida y los precedentes para resolver los casos de alumnos que se hayan visto afectados por circunstancias adversas a fin de compensar, en la medida de lo posible, cualquier desventaja que hayan padecido. A veces, esto conlleva un ligero aumento de la puntuación, y otras veces, la compensación necesaria por no haber realizado un componente completo de examen. En este último caso, puede emplearse un "procedimiento de ausencia de calificación", en el que se utiliza la estadística para calcular la puntuación probable del componente no realizado, basándose en las puntuaciones obtenidas en otros componentes, comparadas con el rendimiento medio de los alumnos en su conjunto. Esta compensación debe aplicarse bajo ciertas condiciones, que aparecen publicadas como parte del reglamento general en el *Vademécum*, el manual de procedimientos que regula el sistema de evaluación del Programa del Diploma.

El comité estudia los casos de administración impropia por parte de colegios que no hayan respetado los plazos o los procedimientos establecidos. Cuando se trata de casos graves que implican una amenaza importante para la seguridad e integridad de los exámenes, o de reiterados casos de administración impropia, puede serle retirada la autorización al colegio en cuestión.

Muchas deliberaciones del comité giran en torno a acusaciones de conducta fraudulenta de alumnos. Dichas acusaciones pueden ser formuladas por el propio colegio, con relación a la conducta de un alumno en un examen, por ejemplo, o pueden formularlas examinadores que crean haber detectado plagio o connivencia en el trabajo de los alumnos. Otro posible tipo de conducta fraudulenta es aquella en la que se incurre cuando los alumnos de una parte del mundo informan del contenido de los exámenes a alumnos de otra parte del mundo antes de que realicen su examen. Esto se hace posible por los diferentes husos horarios, que impiden que puedan realizarse los exámenes simultáneamente en todo el mundo. Se pide a los examinadores que estén atentos para detectar posibles casos de ello. IBO tiende cada vez más a utilizar variantes regionales de los exámenes a fin de reducir las posibilidades de que los alumnos puedan compartir ese tipo de información.

En todos los casos de supuesta conducta fraudulenta, se hacen todas las averiguaciones posibles sobre lo sucedido y se recogen declaraciones del colegio y de los alumnos implicados antes de que el comité tome una decisión final sobre el asunto. A los alumnos que sean considerados culpables de conducta fraudulenta se les deniega la calificación final en esa asignatura y, en consecuencia, no se les puede conceder el diploma.

## 5.8 Publicación de resultados

Los resultados del diploma y de los certificados de las dos convocatorias anuales se publican el 5 de enero y el 5 de julio de cada año para que colegios y universidades puedan acceder a ellos. Se envían electrónicamente, al igual que muchos otros procesos administrativos relativos al sistema de exámenes, como las matriculaciones de alumnos, y la mayoría de registros de puntuaciones. Se les expide a los alumnos una calificación final numérica de 1 a 7 para cada asignatura en la que están matriculados, y aquellos alumnos que siguen el programa completo también reciben notas finales representadas por letras para Teoría del Conocimiento y la Monografía, junto con la puntuación total del Diploma. Puede concederse un diploma bilingüe a los alumnos que:

1. hayan realizado exámenes de dos lenguas A1
2. hayan realizado exámenes de una lengua A1 y de una lengua A2 distinta, o
3. hayan realizado exámenes, o presentado una monografía, en al menos una de las asignaturas del Grupo 3 o del Grupo 4 en una lengua distinta de la correspondiente al examen de Lengua A1.

Generalmente, muy pocos alumnos consiguen la máxima puntuación total posible para el Diploma, que es de 45 puntos. Aproximadamente el 5% de todos los alumnos que cursan el programa completo obtienen más de 40 puntos. El porcentaje de aprobados ha permanecido bastante estable en torno al 80% en los últimos años. A mediados de febrero y mediados de agosto se envía a los colegios la documentación oficial sobre los resultados.

## 5.9 Comentarios y consultas sobre los resultados

En los últimos años, IBO ha puesto mucho énfasis en mejorar y reforzar los comentarios que proporciona a colegios y profesores con relación al sistema de evaluación del Programa del Diploma y a los logros de los alumnos. Esto tiene la doble ventaja de reforzar el apoyo que proporciona la evaluación sumativa a la enseñanza de clase, y de clarificar los mecanismos del sistema de evaluación para los colegios, profesores y alumnos que lo utilizan. Tras cada convocatoria, los exámenes y los esquemas de calificación que los acompañan se ponen a disposición de los colegios para que puedan comprarlos. El equipo examinador supervisor redacta informes de las asignaturas que cubren todos los aspectos generales del rendimiento de los alumnos en cada componente, indicando dónde rindieron más y dónde menos, y haciendo recomendaciones para mejorar su preparación. En los informes también figuran las bandas de calificaciones finales que se aplican a cada componente. Los informes de las asignaturas se ponen a disposición directa de los profesores a través del centro pedagógico en línea (CPEL), un sitio web dedicado a proporcionar apoyo profesional a los profesores del BI.

Poco después de la publicación de los resultados oficiales, se pone automáticamente a disposición de los colegios información electrónica sobre la puntuación moderada y la nota final correspondiente a cada componente de la evaluación de cada alumno. Las puntuaciones y las notas finales de los componentes no forman parte de los resultados que se publican oficialmente, por los motivos señalados en la sección 3.2. Se proporcionan a los colegios como información complementaria útil, que indica los puntos relativamente fuertes y débiles de sus alumnos en los diferentes componentes del modelo de evaluación de una asignatura. También se informa a los colegios de cómo influyó la moderación en las puntuaciones concedidas por sus profesores en cada componente de la evaluación interna. Además, se envía electrónicamente a los colegios un formulario de comentarios sobre la evaluación interna, preparado por el moderador, informando brevemente sobre los puntos fuertes de la evaluación interna de cada asignatura y lo que puede mejorarse.

Además de estos tipos más generales de comentarios, los colegios pueden utilizar el servicio de consultas sobre los resultados para hacer un seguimiento de casos más concretos. Existen tres categorías de consultas por las cuales se debe abonar una tasa. La categoría 1 consiste en volver a corregir el trabajo que un alumno haya realizado en una asignatura y que haya sido evaluado externamente, si el colegio o el alumno creen que el resultado no refleja con justicia su rendimiento. Si se mejora la calificación de la asignatura como resultado de esta nueva puntuación, no se cobrará la tasa. No se bajan las calificaciones finales de la asignatura cuando se vuelve a calificar el trabajo, aunque puede que sí se bajen las de los componentes.

La categoría 2 permite a los colegios solicitar la devolución de las copias del trabajo de un grupo completo de alumnos de un colegio correspondiente a un componente determinado evaluado externamente, pudiendo tratarse de un componente de examen o no de examen. Los profesores pueden así ver cómo se calificó un trabajo de alumnos de su clase. Se anima a los examinadores a escribir comentarios breves y constructivos cuando lo puntúen. Éstos ayudan a los moderadores y examinadores supervisores que puedan tener que revisar las puntuaciones posteriormente, y también proporcionan más comentarios útiles para los profesores. Aunque el propósito de esta categoría de consulta es proporcionar comentarios que permitan a los profesores orientar su enseñanza en el futuro, también es posible que tras dicha consulta un colegio solicite una revisión de la nota (informe de categoría 1).

La categoría 3 consiste en un informe redactado por el moderador, explicando con mucho más detalle del que es posible en el formulario de comentarios de la evaluación interna mencionado antes, por qué las puntuaciones de la evaluación interna realizada por un profesor, correspondientes a cada criterio, se confirmaron o ajustaron. También se facilita información más detallada sobre la adecuación de las tareas que, en efecto, se utilizaron para la evaluación interna. Se pretende que los colegios utilicen este servicio cuando hayan recibido ya el formulario de comentarios sobre la evaluación interna y los datos de ajuste de las puntuaciones moderadas, si creen que necesitan más información para comprender por qué se ha aplicado un determinado ajuste de moderación.

Todas las fuentes de comentarios descritas en esta sección se ponen a disposición de profesores y colegios con el objetivo expreso de utilizar el sistema de evaluación formal del Programa del Diploma para hacer más eficaz la enseñanza de clase y para mejorar el aprendizaje del alumno.

## Apéndice A

# Validez, fiabilidad y generalizabilidad: información adicional

Esta sección va destinada a aquellos lectores que necesiten una descripción más detallada de dos conceptos fundamentales de la evaluación: la validez y la fiabilidad.

### A.1 Validez

La obra clásica sobre validez es la de Messick (1989). Él reconoció la necesidad de considerar la validez como un concepto unitario y la validez de constructo como el tema subyacente. La validez de constructo se refiere al grado en que un instrumento de evaluación refleja adecuadamente una capacidad o destreza, o ámbito subyacente, de modo que ningún elemento significativo del constructo deje de estar representado adecuadamente, y no se incluyan variables adicionales ajenas al constructo. Todas las tareas de un instrumento de evaluación deben dirigirse única y exclusivamente al constructo subyacente. Es probable que una visión psicométrica de un constructo, por ejemplo, la expresión escrita, sea mucho más estrecha que la visión de la evaluación del rendimiento de ese mismo constructo, que probablemente pondrá un mayor énfasis en la creatividad y las capacidades productivas complejas. En la evaluación del rendimiento, no es probable que puedan representarse todas las facetas de un constructo dado en la evaluación misma, y será necesario un muestreo del contenido del ámbito. En general, la validez de constructo se juzga cualitativamente mediante la opinión de expertos.

En opinión de Messick, si el constructo es válido, es probable que también se den los otros tipos de validez. La validez predictiva se refiere a la capacidad de un instrumento de evaluación para prever el rendimiento de futuras evaluaciones de naturaleza similar. La validez concurrente se refiere al grado de correlación entre el rendimiento en un instrumento de evaluación y el rendimiento en otro instrumento de evaluación diseñado para medir el mismo constructo. Está claro que la validez concurrente y la validez predictiva tienen mucho en común, y pueden medirse estadísticamente por correlación. Si existe un constructo compartido, representado adecuadamente en las diferentes formas del instrumento de evaluación, parecería lógico esperar que la validez concurrente y la predictiva fueran altas. Sin embargo, pueden surgir problemas con la validez predictiva cuando se produce una importante selección en el progreso del alumnado de un nivel de evaluación a otro futuro, cuando tienen lugar experiencias educativas distintas entre un nivel de evaluación y el siguiente, y cuando se desarrolla o elabora el constructo entre un nivel de evaluación y el siguiente. La validez de contenido, como su nombre indica, se refiere al contenido real de conocimientos y destrezas de cada elemento/tarea de evaluación, y si es apropiada para el constructo en cuestión. La validez de contenido se mide generalmente mediante la opinión profesional, y obviamente existe un importante solapamiento entre ésta y la validez de constructo.

Messick también argumentó que la validez de las evaluaciones depende en parte de la interpretación de los resultados de la evaluación y de las consecuencias deseadas y no deseadas de éstos. Si se da a los resultados de la evaluación un uso inadecuado, se reduce entonces la validez de ésta, lo que también implica el fracaso de la validez de constructo de la evaluación. Moss (1992, citado en Gipps, 1994) afirmó que “el propósito esencial de la validez de constructo es justificar una interpretación concreta de una prueba explicando el comportamiento que el resultado de la prueba resume”. Esto supone un cambio de dirección importante con respecto a la práctica psicométrica tradicional que considera la validez únicamente en términos de constructo y de contenido, junto con estudios de correlación de rendimiento en pruebas destinadas a medir el mismo constructo. La noción relativamente reciente de validez consecucional tiene implicaciones evidentes en lo que respecta al papel del “efecto de repercusión” en la enseñanza, y ha servido para

impulsar el desarrollo de la evaluación del rendimiento en ciertas partes del mundo. El reconocimiento de que la evaluación “de alto riesgo” tiende a tener una gran influencia estructural en los sistemas educativos (por ejemplo, Frederiksen y Collins, 1989; Broadfoot, 1996), junto con el reciente y creciente impacto de la legislación sobre pruebas en EE.UU., han contribuido a la importancia que ahora se le da a las consecuencias sociales de la evaluación en los estudios sobre validez.

## A.2 Fiabilidad

Existen dos enfoques principales en la medición de la fiabilidad con relación al instrumento de evaluación: el intrínseco y el extrínseco. El intrínseco implica considerar las propiedades del instrumento de evaluación en sí mismo. El test-retest y las formas paralelas son métodos intrínsecos, que se ocupan de evaluar la estabilidad de las respuestas del alumno. En un procedimiento de test-retest, se les da a los alumnos el mismo test dos veces, separados por un breve intervalo de tiempo y sin que los alumnos sepan los resultados, para ver hasta qué punto responden lo mismo en ambas ocasiones. En los tests psicométricos no puede darse por hecho que las respuestas vayan a ser muy similares, porque muchos de los ítems del test están diseñados para operar en el máximo nivel de incertidumbre de los alumnos (es decir, para producir el máximo grado de diferenciación entre alumnos). En la evaluación del rendimiento es menos aplicable este procedimiento, porque el rendimiento de los alumnos la segunda vez probablemente dependerá mucho de la forma en la que respondieron la primera vez, inflando artificialmente la medición. El método de las formas paralelas estudia el grado de semejanza de las respuestas a dos versiones distintas del “mismo” test. En los tests psicométricos, esto implica a menudo la preparación de otro test partiendo de un banco común de ítems de test, que se les da a los mismos alumnos. Deben tenerse en cuenta las estadísticas de rendimiento de los ítems ligeramente distintos de las muestras específicas de ítems que componen cada test.

Tomando los exámenes como ejemplo típico de evaluación del rendimiento, los exámenes de la misma asignatura que se preparan para distintas convocatorias pueden constituir formas paralelas, aunque debe tenerse en cuenta que no se espera que los mismos niveles de rendimiento se recompensen con las mismas puntuaciones en las dos convocatorias. La equiparación debe hacerse a un nivel más alto de resultados, ya sean notas finales o puntuaciones que, de algún modo, hayan sido procesadas. Aunque los métodos de test-retest y de las formas paralelas para calcular la fiabilidad son sólidos y proporcionan datos de mucha calidad, no se utilizan corrientemente debido a las exigencias prácticas y al gran volumen de recursos que implican (Feldt y Brennan, 1989). Una medida intrínseca de fiabilidad mucho más accesible y ampliamente utilizada se basa en comparar el rendimiento del alumno en una parte de un test con el rendimiento en el resto de éste, en otras palabras, en estudiar la coherencia interna de un test. Se considera adecuado porque todos los ítems (preguntas) del test están diseñados para evaluar el mismo atributo o destreza: sólo se incorpora un ítem al test si existe una buena correlación entre el rendimiento de los alumnos en ese ítem y su rendimiento en el test en su conjunto. Esto requiere que un solo grupo de alumnos haga el test una sola vez.

El test se puede dividir de varias formas: comparando la primera mitad con la segunda mitad del mismo o los ítems pares con los impares (método de las dos mitades), o mediante fraccionamientos más sofisticados, combinando estadísticamente todas las formas posibles de dividir el test en dos. En efecto, las dos mitades de un único test se consideran formas paralelas con una deficiencia: que las evaluaciones no están separadas en el tiempo.

Estas medidas de coherencia interna se han convertido en las más utilizadas para medir la fiabilidad de los tests estandarizados. Desgraciadamente, la coherencia interna no es en modo alguno una característica que pueda esperarse de la mayoría de las evaluaciones del rendimiento, tales como los exámenes. No se espera que el rendimiento del alumno en una tarea deba ser igual a su rendimiento en otra tarea en el mismo instrumento de evaluación, aunque exista en la práctica a menudo una relación bastante estrecha entre ambos. De hecho, puede muy bien haberse diseñado la tarea deliberadamente para evaluar el rendimiento en tipos muy diferentes de destrezas en un ámbito muy extenso del área temática que se examina. Como consecuencia, no tiene mucho sentido aplicar mediciones de coherencia interna de fiabilidad a lo que no sean tests estandarizados. Es mucho más que una curiosidad señalar que, cuando se han aplicado al

mismo test diferentes tipos de mediciones de fiabilidad (test-retest, formas paralelas y coherencia interna), se han encontrado a menudo importantes discrepancias entre las diferentes mediciones. Por ejemplo, un estudio del test de Iowa de destrezas básicas (ITBS, 1986, citado en Wood, 1991) descubrió que los cambios en el rendimiento de los alumnos en diferentes días ofrecían una medida de fiabilidad muy reducida en comparación con la que ofrecía la coherencia interna.

### A.3 Fiabilidad del evaluador

El otro aspecto de la fiabilidad, el aspecto extrínseco de la coherencia de la corrección es, sin embargo, un área de la evaluación del rendimiento que causa gran preocupación. Los tests objetivos estandarizados no permiten ninguna variación a la hora de calificar diferentes respuestas: cada una es claramente correcta o incorrecta. No obstante, las evaluaciones del rendimiento, que dependen mucho de la opinión del evaluador, son vulnerables a las diferencias de valoración de los méritos de un determinado trabajo del alumno. Estas preocupaciones son de dos tipos: la fiabilidad intraevaluador (qué grado de coherencia tiene un evaluador determinado al conceder una puntuación al mismo trabajo en distintas ocasiones) y la fiabilidad interevaluador (qué grado de semejanza tienen las puntuaciones concedidas a un mismo trabajo por dos evaluadores diferentes).

Se han publicado pocos trabajos de investigación recientemente sobre la fiabilidad del proceso de corrección en los sistemas de exámenes que dependen de la opinión del examinador. Sin embargo, investigaciones anteriores descubrieron que existía un bajo nivel de fiabilidad interevaluador e intraevaluador (University of Cambridge Local Examinations Syndicate, 1976; Willmott y Nuttall, 1975; Nuttall y Willmott, 1972; Murphy, 1978 y 1982). La investigación demuestra que los dos niveles de fiabilidad se ven afectados por la naturaleza de la tarea de evaluación de que se trate. Las preguntas de respuesta corta y las respuestas muy estructuradas a preguntas analíticas pueden calificarse con mayor fiabilidad que las tareas de redacción abiertas o las interpretaciones/productos artísticos. Se ha demostrado que la preparación de los examinadores y el proporcionarles esquemas de calificación elevan la fiabilidad de la calificación a niveles altos (Brown, 1992; Shavelson et ál., 1992).

La cuestión de la fiabilidad es la que fundamentalmente separa la evaluación psicométrica de la del rendimiento. Es fácil lograr un alto grado de fiabilidad o, al menos, de coherencia interna en los tests psicométricos, lo que tiene mucho valor en la comercialización de un test y tiene, además, valor a largo plazo, por la larga vida útil de los tests estandarizados. En cuanto a los instrumentos de evaluación del rendimiento, no se persigue como resultado la coherencia interna, y se pone el énfasis en la fiabilidad de la corrección. Sin embargo, tiene poco sentido hacer grandes esfuerzos por establecer el grado de fiabilidad del evaluador para un determinado instrumento de evaluación, como un examen, que pierde su valor después de ser utilizado una vez. En lugar de ello, se pone el énfasis en los mecanismos de control de calidad de los procesos de corrección de todo el sistema, a fin de elevar la coherencia del evaluador a los niveles más altos posibles para cada construcción distinta de un determinado instrumento de evaluación.

Estos mecanismos de control de calidad son una característica importante de todos los sistemas de evaluación del rendimiento, ya sea en EE.UU., Europa, Australasia u otras partes del mundo. En algunos sistemas, se reúne a todos los examinadores en un lugar para que lleven a cabo la corrección bajo la vigilancia de un examinador con experiencia que comprueba constantemente su trabajo. En ocasiones, puede mejorarse la fiabilidad aún más pidiendo a cada examinador que trabaje solamente en una pregunta o tarea del instrumento de evaluación. Alternativamente, los examinadores pueden trabajar individualmente, pero enviar una muestra de lo que han corregido al supervisor para que lo vuelva a calificar. Basándose en la comparación del conjunto de pares de puntuaciones, puede aplicarse un ajuste estadístico (o ajuste de moderación) a las puntuaciones iniciales del examinador. Un tercer enfoque consiste en que el trabajo del alumno sea corregido independientemente por dos examinadores; si están de acuerdo dentro de ciertos límites se concede la puntuación media de los dos, pero si existe una gran discrepancia, el trabajo se corrige una tercera vez por un supervisor que actúa de árbitro. En cuarto lugar, en algunos sistemas, se asigna a los examinadores el trabajo de los alumnos deliberadamente al azar, en el entendimiento de que las distribuciones de puntos resultantes serán muy similares en el caso de todos los examinadores. Se aplican

ajustes estadísticos a la distribución de puntos de cada examinador para que coincida con la obtenida por el examinador supervisor con más experiencia del equipo.

Estos procesos de control de calidad precisan recursos distintos, y tienen una viabilidad diferente según el contexto. También varían en cuanto a sus relativas ventajas y desventajas, pero todos tienen como objetivo resolver el problema de garantizar niveles satisfactorios de fiabilidad interevaluador e intraevaluador.

## **A.4 Sistemas de calificación y fiabilidad**

Una característica de muchos sistemas de evaluación del rendimiento en el ciclo superior de la enseñanza secundaria es que están compuestos por varios instrumentos o componentes de evaluación distintos, cada uno diseñado para evaluar los conjuntos de conocimientos y destrezas o capacidades, a veces bastantes distintos, que cubre un curso. Esto es particularmente cierto de la evaluación del rendimiento “de alto riesgo”. Utilizar en diferentes ocasiones varios instrumentos de evaluación que se evalúan independientemente reduce el impacto de la variabilidad del evaluador y del alumno, además de ofrecer la oportunidad de componer una imagen de los logros del alumno en diferentes contextos. Asimismo, los resultados de estas evaluaciones del rendimiento combinadas son mucho más reducidos, por ejemplo: dos posibles resultados (aprobado/no aprobado); o no aprobado, aprobado, notable y sobresaliente; o un sistema de calificación que comprenda un número restringido de éstas, por ejemplo, de A a G o de 1 a 9.

El motivo de este enfoque es que la fiabilidad del resultado de un alumno dentro de una gama restringida de posibles calificaciones y basado en la combinación de varios instrumentos de evaluación será muy superior a la fiabilidad de la puntuación que se conceda en un único instrumento de evaluación. Es inevitable la pérdida de información que se produce al reducir la escala de resultados del alumno de, por ejemplo, 400 puntos a, por ejemplo, cinco calificaciones. Sin embargo, cuando se dan los resultados de esas calificaciones, su fiabilidad puede ser tan elevada como la de los resultados de las puntuaciones de un test objetivo estandarizado. El que se reduzca el margen de posibilidades de establecer diferencias entre alumnos al disminuir la escala de puntuación de 400 puntos a 5 puntos no constituye una preocupación importante, porque establecer diferencias precisas entre alumnos no constituye un objetivo primordial de la evaluación del rendimiento.

Entre los dispositivos adicionales para mejorar la fiabilidad del resultado final en un sistema de evaluación del rendimiento pueden incluirse más comprobaciones por parte de los examinadores supervisores con mayor experiencia, y volver a calificar el trabajo de los alumnos cuya puntuación combinada final los sitúe cerca del límite entre una calificación final y la siguiente. En tales casos, un pequeño error de medición puede tener un impacto decisivo en el resultado final y en las oportunidades del alumno en el futuro en un entorno “de alto riesgo”. Cuanto más reducido sea el número de notas finales, mayor será el impacto que pueda tener para un alumno la nota equivocada. Además, en algunos sistemas de evaluación, los alumnos pueden solicitar que se vuelva a calificar su trabajo después de que se hayan expedido los resultados si creen que el resultado obtenido no refleja con justicia sus logros.

En el contexto de los sistemas de calificación, existe otra cuestión de fiabilidad aplicable a algunos sistemas de evaluación del rendimiento: la fiabilidad de las notas finales. Hay algunos instrumentos de evaluación del rendimiento, por ejemplo, los exámenes, cuya dificultad variará ligeramente en cada versión que se prepara para cada vez, a pesar de los esfuerzos coordinados para conseguir que todas las versiones sean equiparables en cuanto a su dificultad. Dado que las notas que se conceden al final de la evaluación tienen como objetivo representar estándares de logro coherentes, aunque puede ser ligeramente más fácil o más difícil obtener puntos en diferentes ejemplos de los instrumentos de evaluación, se sigue que la escala para convertir los puntos en notas finales puede precisar ajustes cada vez. Esto se hace estableciendo bandas que determinan la puntuación mínima requerida para conseguir cada nota final. En algunos sistemas, estas bandas se determinan siguiendo la opinión de expertos en la asignatura, que revisan el conjunto de los trabajos presentados por los alumnos en cada ocasión. Para mejorar la fiabilidad de las notas finales, se establecen los límites entre una y otra mediante el consenso de opiniones de un grupo de examinadores con experiencia, que tienen ejemplos anteriores de trabajo con bandas de calificación para que les sirvan

de referencia, y datos estadísticos para apoyar su decisión. Sin embargo, debido a la naturaleza personal y subjetiva de la contribución de cada experto a la opinión definitiva, existirá inevitablemente cierta variabilidad en las opiniones consensuadas (Cresswell, 1996).

A veces se adopta un enfoque alternativo que consiste en ajustar las distribuciones de notas finales de los alumnos en una determinada asignatura, para que coincidan con la distribución de esos mismos alumnos en otras asignaturas. Se trata de un proceso estadístico circular e iterativo que sólo puede aplicarse en el contexto de una distribución general de notas establecida en todas las asignaturas. Este enfoque es un ejemplo de cómo una distribución de notas predeterminada, algo normalmente asociado a las pruebas estandarizadas, puede introducirse en la evaluación del rendimiento. En un sistema así, no puede interpretarse que las calificaciones representan estándares establecidos de logro, sino que indican, en lugar de ello, la posición de un alumno en ese grupo concreto. Dado que no existe intención expresa de conceder la misma nota final para el mismo nivel de logro en diferentes casos en la evaluación, el concepto de fiabilidad del proceso de calificación final no se aplica en este sistema, aunque la fiabilidad de la corrección continúa siendo, naturalmente, importante.

## A.5 Generalizabilidad

El valor de una puntuación o nota final concedida como resultado de un sistema de evaluación reside en su generalizabilidad: ¿cuánto nos dice el resultado sobre lo que el alumno podría hacer otras veces, en contextos distintos pero relacionados? Toda evaluación sólo puede basarse en una muestra de conducta posible, con el propósito de que el logro alcanzado en esa muestra pueda ser de aplicación general al logro en todo el universo de un campo concreto de conducta (Nuttall, 1987). Para que pueda darse la generalización, debe definirse todo el campo de conducta (o constructo), y la evaluación debe ser fiable. La generalizabilidad, por tanto, depende tanto de la validez como de la fiabilidad, y conecta a ambas.

Uno de los mayores desafíos para la generalizabilidad es el papel del contexto. Se da aquí un paralelismo entre evaluación y aprendizaje. Desde una perspectiva tradicional del aprendizaje, puede haberse supuesto en el pasado que resultaba más rentable presentar los conceptos complejos, aplicables a muchas áreas, de forma descontextualizada, abstracta y descompuesta en distintas partes. Sin embargo, un modo más reciente de entender cómo se desarrolla el aprendizaje (por ejemplo, Wood, 1998; Murphy, 1999; Shepard, 1992), pone el énfasis en la compleja naturaleza de éste, en la necesidad de conectar estrechamente los procesos abstractos con problemas y actividades concretas, y en la naturaleza esencialmente social del aprendizaje.

El enfoque del aprendizaje así situado influirá inevitablemente tanto en la forma de realizar la evaluación, en sí una actividad social, como en lo que se infiera de ella, si se quiere mantener una relación de apoyo mutuo entre evaluación y aprendizaje. En cuanto a los tests objetivos estandarizados, debe establecerse la extensión del contexto al que se aplica el test. En cuanto a la evaluación del rendimiento, que generalmente se basa en una perspectiva más amplia del constructo subyacente, la forma más eficaz de aumentar la generalizabilidad es aumentar el número y la distribución de los diferentes tipos de tarea en la evaluación global (Linn, 1993). Linn et ál. (1991) también han indicado que, al igual que el concepto de validez se ha ampliado para abarcar las consecuencias de la evaluación (véase el apéndice A.1), también hay argumentos para defender que el concepto de fiabilidad deba ampliarse para abarcar lo que pueda inferirse de la puntuación de un test con respecto a un ámbito más amplio de logro. Si se considerase la fiabilidad de este modo, las medidas de coherencia interna, tan utilizadas para los tests de opción múltiple estandarizados, serían inadecuadas. En el contexto de los tests de opción múltiple estandarizados, la fiabilidad y la validez se centran internamente en el test mismo, mientras que en la evaluación del rendimiento ambos conceptos tienen una perspectiva enfocada más hacia el exterior, centrándose en la utilidad de la evaluación y su efecto en otros aspectos de la educación. En términos generales, es probable que sean más generalizables (o menos específicos del contexto) los resultados de la evaluación del rendimiento que los resultados de las pruebas psicométricas.

## Apéndice B

### Política de evaluación del Programa del Diploma

1. Toda evaluación que se lleve a cabo en las asignaturas del Programa del Diploma debe estar directamente relacionada con el curso y sus objetivos, y debe administrarse, en la medida que sea posible en la práctica, a través de una política de pruebas discretas en cada entorno de evaluación (pruebas escritas, evaluación interna, etc.). Debe utilizarse una gama completa de técnicas de evaluación que reflejen la vocación internacional de IBO. Debe aplicarse la misma metodología de evaluación a asignaturas conexas, pero toda diferencia sustancial en la naturaleza del Nivel Superior y del Nivel Medio de una asignatura debe verse reflejada en sus modelos de evaluación respectivos.
2. Los procesos de evaluación y concesión de calificaciones finales del Programa del Diploma deben garantizar la igualdad de trato para todos los alumnos, con independencia del colegio, la asignatura, la lengua de respuesta y la convocatoria. Todos los procesos de evaluación y la concesión de calificaciones finales deben basarse en datos objetivos y no deben verse sometidos a ningún tipo de sesgo.
3. Todos los cursos deben tener normalmente tres o cuatro componentes de evaluación distintos. Cuando proceda, incluirán componentes de evaluación interna (evaluados en el colegio) y componentes de evaluación externa. Ningún componente de evaluación debe por sí solo representar menos del 20% o más del 50% del total de la evaluación, y todos los componentes evaluados internamente no deben contribuir juntos en más del 50% a dicho total. Debe existir un equilibrio tal entre la evaluación interna y la evaluación externa que garantice que todos los objetivos del curso se evalúen adecuada y debidamente.
4. La duración de los exámenes escritos no debe ser superior a cinco horas en total en el Nivel Superior, ni a tres horas en el Nivel Medio. Ninguna de las pruebas debe durar en sí misma más de tres horas. Siempre que sea posible y no se vea comprometida la fiabilidad y validez de los exámenes, su duración será inferior al máximo prescrito. Esta restricción es especialmente pertinente con respecto a aquellas asignaturas en las que la evaluación interna y otros componentes calificados externamente formen una parte importante del modelo de evaluación de la asignatura.
5. Las puntuaciones de profesores y examinadores se deben moderar conforme a un modelo de corrección/revisión de la corrección seguido de una comparación estadística que genera una ecuación de moderación. No se realiza la moderación entre componentes. La revisión de la corrección se debe basar en los mismos criterios de evaluación utilizados en la corrección inicial y en muestras de trabajo enviadas a un examinador que actuará en calidad de moderador.
6. En la evaluación interna deben abordarse primordialmente aquellas capacidades o destrezas y áreas de conocimiento que es menos adecuado abordar a través de exámenes externos; no debe tratarse como otro medio en el que los alumnos demuestren, en un contexto diferente, lo que también pueden demostrar en un examen. No debe duplicarse injustificadamente la evaluación de las mismas destrezas o capacidades en la evaluación interna y en los exámenes externos.
7. La evaluación interna no debe utilizarse como instrumento para comprobar si se está cubriendo el contenido de la asignatura, sino que debe centrarse en evaluar si el alumno está aprendiendo ciertas destrezas concretas. Cuando sea necesario, la extensión con la que se ha cubierto el contenido de la asignatura deberá evaluarse en exámenes externos.
8. Las tareas de evaluación interna no deben duplicar el tipo de trabajo que se lleve a cabo en las monografías correspondientes a la misma asignatura.

9. Siempre que sea posible, las tareas de evaluación interna deben constituir una parte integral de la enseñanza normal de clase (o de los deberes para casa) correspondientes a esa asignatura. No debe tratarse de actividades adicionales. El trabajo que se lleva a cabo en la evaluación interna debe ser parte de la experiencia de aprendizaje de cada alumno.
10. Para que las puntuaciones de la evaluación interna puedan contribuir de forma fiable a la calificación final del alumno en una asignatura, el trabajo que contribuye al menos a la mitad de la puntuación de la evaluación interna debe ser susceptible de moderación. Este es el mínimo establecido y es preferible, siempre que sea posible, que todo el trabajo que contribuye a la puntuación de la evaluación interna pueda ser moderado.
11. Cuando, para la evaluación interna de una asignatura del Programa del Diploma, se realicen distintas tareas durante un período prolongado de tiempo (por ejemplo, para preparar un portafolio de trabajo) debe considerarse la posibilidad de que el rendimiento del alumno mejore durante este período. Por tanto, la puntuación final de la evaluación interna debe reflejar el mejor nivel de rendimiento del alumno durante el curso, y no ser simplemente una media de su rendimiento durante todo el curso.
12. Aunque la evaluación interna puede contribuir entre un 20% y un 50% al resultado de cualquiera de las asignaturas, los valores más altos de esta gama sólo deben utilizarse cuando haya motivos concretos para darle una ponderación alta al trabajo evaluado internamente.
13. El trabajo evaluado internamente debe producirse en condiciones debidamente documentadas y ser comunes a todos los colegios para cada curso. En particular, debe describirse completamente el papel del trabajo en equipo, el grado de ayuda que pueden prestar los profesores, hasta qué punto pueden los alumnos utilizar recursos externos, y cuántas veces puede volver a redactarse el trabajo.
14. La cantidad de trabajo evaluado internamente que se especifique para un curso concreto no debe superar el mínimo necesario para cumplir los objetivos de dicho curso. Deben establecerse, siempre que sea posible, límites en el número de palabras de las tareas evaluadas internamente. El límite máximo de palabras no debe superar al necesario para completar la tarea.

## Referencias

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

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. [disponible en español: *Taxonomía de los objetivos de la educación: Clasificación de las metas educativas*. Tomo 1. Alicante: Ed. Marfil S.A. Alcoy, 1979]

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* (3<sup>rd</sup> 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. [disponible en español: *Estructuras de la mente: la teoría de las inteligencias múltiples*. México: FCE, 1983]
- 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*, 2<sup>nd</sup> 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: John 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, 135–144.
- Harlen, W (ed) (1994) *Enhancing Quality in Assessment*, BERA Policy Task Group on Assessment, Paul Chapman Publishers.
- Hieronimus, 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.
- Organización del Bachillerato Internacional (1999), Programa del Diploma, *Guía de Lengua A1*, Cardiff: IBO.
- Organización del Bachillerato Internacional (2001a), Programa del Diploma, *Guía de Historia*, Cardiff: IBO.
- Organización del Bachillerato Internacional (2001b), Programa del Diploma, *Guía de Biología*, Cardiff: IBO.
- Organización del Bachillerato Internacional (2003a), *Perceptions of the International Baccalaureate Diploma Programme*, Cardiff: IBO.
- Organización del Bachillerato Internacional (2003b), *Probidad académica: información para los colegios*, 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* (3<sup>rd</sup> 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*, (2<sup>nd</sup> 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. [disponible en español: *Evaluación basada en criterios*. Madrid: Ed. Magisterio Español S. A., 1983]
- 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*, Wiley: New York. [disponible en español: *Pensamiento y lenguaje*. Buenos Aires: La pléyade, 1977]
- 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.
- William, 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* (2<sup>nd</sup> 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: University Press.









