

Enfoques estadísticos del establecimiento de límites de calificación

Resumen ejecutivo

Octubre de 2022

Resumen ejecutivo

Contexto

El Bachillerato Internacional (IB) es una fundación internacional sin ánimo de lucro que ofrece una serie de programas educativos para el alumnado de 3 a 19 años. Estos programas educativos sirven de alternativa a los programas propios de cada país, con sus currículos y evaluaciones. Como resultado, una de las muchas tareas que tiene que afrontar el IB para desarrollar esos programas es el establecimiento y mantenimiento de un estándar para esas evaluaciones, con el fin de garantizar la mayor equidad y comparabilidad de un año a otro.

Sobre este estudio

Este estudio se centra en un elemento del modo en que el IB convierte las puntuaciones de sus evaluaciones en calificaciones y mantiene los estándares de un año a otro. Más concretamente, se concentra en los distintos enfoques que son posibles para crear límites recomendados estadísticamente para el Programa del Diploma (PD), el Programa de Orientación Profesional (POP) y el Programa de los Años Intermedios (PAI) en sus puntos fuertes y débiles, y en las situaciones en las que se debe preferir un enfoque a los demás. En este estudio se definen los límites recomendados estadísticamente como las estimaciones para los límites de calificación (la puntuación mínima que se debe alcanzar para obtener una calificación) basadas en una comparación estadística de los puntos fuertes relativos del grupo de un año en comparación con el del año anterior (IB, 2018).

Con este fin en mente, el estudio se centró en:

- Revisar la bibliografía especializada reciente y más significativa acerca de los diferentes enfoques para calcular los límites recomendados estadísticamente, incluida una revisión de sus puntos fuertes y débiles y de cuándo es más probable que sean adecuados. Además, la revisión incluyó también otros enfoques estadísticos para el establecimiento de límites de calificación cuando los límites recomendados estadísticamente no sean adecuados, como cuando no haya un grupo anterior que se evalúe.
- Documentar una serie de diferentes contextos para el establecimiento de límites de calificación que sean pertinentes para el IB, basados en elementos como el tamaño del grupo de estudiantes, la tasa de crecimiento, etc. para cada asignatura.
- Analizar la precisión e idoneidad de cada enfoque para calcular los límites recomendados estadísticamente en una serie de contextos diferentes que han sido identificados como pertinentes para el IB.

Revisión de la bibliografía especializada

La revisión de la bibliografía especializada constituyó la primera fase del proyecto y su objetivo era revisar y mejorar los procedimientos de establecimiento de los límites recomendados estadísticamente del IB. Idealmente, el objetivo último del proyecto era que los límites recomendados estadísticamente proporcionaran una estimación precisa de la ubicación donde deben estar los límites de calificación para que rara vez se requiera un ajuste (o, al menos, ajustes menores que los límites recomendados estadísticamente en vigor).

Teniendo esto en cuenta, esta revisión de la bibliografía especializada pretende alcanzar los siguientes objetivos:

1. Presentar el “universo” de procedimientos estadísticos de fijación de estándares, incluidos:
 - a. Los requisitos para posibilitar su uso
 - b. Las ventajas y desventajas relacionadas con otros enfoques
2. Establecer valoraciones iniciales sobre qué procedimientos podrían ser más adecuados o inadecuados para cada contexto del IB

Se revisó una abundante bibliografía especializada para recoger información sobre las metodologías estadísticas en uso para la fijación de estándares. A grandes rasgos, se dividen en dos categorías: la equiparación de puntuaciones y los enfoques basados en la predicción. Resumimos a continuación las técnicas que se cubren en ambas categorías, entre otras:

- Equiparación de puntuaciones
 - Técnicas básicas de equiparación (media, lineal, equipercantil, etc.)
 - Técnicas de suavizado
 - Diseños de grupos no equivalentes
 - La Teoría de la Respuesta al Ítem (TRI)
- Basadas en la predicción
 - Formas de deducir una predicción
 - Indicadores externos de diferencias entre los grupos
- Formas de combinar varios enfoques

Basándonos en ello, hemos extraído algunas conclusiones iniciales acerca de los enfoques de la fijación de estándares que es probable que sean viables (o no) para cada contexto del IB. Los enfoques basados en los logros anteriores parecen inviables debido a la falta de esa información, mientras que los diseños de grupos no equivalentes tampoco son prácticos ya que los elementos ancla pondrían en peligro la seguridad de las evaluaciones del IB.

Esto nos deja, para generalizar, tres enfoques amplios que parecen prometedores para el IB:

- a. Técnicas básicas de equiparación
- b. Enfoques simultáneos basados en los logros
- c. Enfoques que intentan mantener el resultado anterior (es decir, a través de centros comunes)

En conjunto, las técnicas básicas de equiparación son adecuadas en situaciones donde los dos grupos de estudiantes son comparables en capacidad. Sin embargo, solo cerca de la mitad de los contextos del IB cumplen ese supuesto. Es preciso recordar que los enfoques básicos de equiparación se pueden aplicar prácticamente en cualquier circunstancia (solo precisan una muestra de pequeño tamaño), lo que podría significar que en algunos casos son la única opción viable. La cuestión es si es aconsejable su uso (en otras palabras, si es probable que los grupos sean distintos) o si sería preferible apoyarse únicamente en enfoques basados en juicios de valor.

Los enfoques simultáneos de equiparación, como el método del “Resumen instantáneo de logros sin calificaciones” (ISAWG, por sus siglas en inglés) desarrollado por Benton en 2017, son potentes y son adecuados para los programas del IB debido a que incluyen un amplio rango de asignaturas. Además, proporcionan el enfoque de equiparación más convincente (con cierto margen) para algunos de los contextos más difíciles, incluidas asignaturas muy pequeñas, aquellas con cambios drásticos en los grupos o asignaturas totalmente nuevas. Sin embargo, los enfoques ISAWG son extremadamente complejos y presentan un gran número de opciones y modificaciones disponibles (incluso cuando se comparan con los demás enfoques de este artículo). Parece probable que ISAWG sea un método que puede ofrecer soluciones para los contextos más difíciles del IB, pero requeriría una cantidad importante de esfuerzo para probarlo e implementarlo adecuadamente —esfuerzo que podría ser desproporcionado para los beneficios que ofrece—. El enfoque también tiene otras desventajas, ya que es complicado implementarlo y es una “caja negra” en términos de facilidad para ser explicado a personas no familiarizadas con él.

Los enfoques que intentan mantener el resultado anterior para un subconjunto del grupo de estudiantes (como el enfoque de “centros comunes”) son un modo bien establecido de intentar explicar los cambios en los grupos, que es viable siempre que el grupo tenga el tamaño suficiente y que haya suficientes centros que cursen la asignatura de un año a otro. Aunque no sea un método tan sólido como los logros anteriores para mantener los resultados, sigue siendo superior a muchos otros enfoques, ya que intenta explicar la capacidad de los grupos a lo largo del tiempo. También es apropiado en casi todos los contextos del IB, con la excepción de los grupos muy pequeños y las asignaturas completamente nuevas (aunque es posible emplear los centros comunes para vincularlo con una asignatura similar o ya existente, por muy dudoso que parezca).

Las fases posteriores del proyecto pueden recurrir a esta revisión para determinar los enfoques que merezca la pena seguir implementando para servir de modelo y evaluar su idoneidad en los distintos contextos de calificación del IB.

Análisis de la simulación de los límites recomendados estadísticamente

Este informe resume los hallazgos del análisis y modelización que incluyen la fase final de un proyecto que intenta revisar los procedimientos del IB para establecer límites recomendados estadísticamente, con el objetivo de facilitar su perfeccionamiento y mejora. Teniendo esto en cuenta, este sería un breve resumen de los objetivos generales de dicho análisis:

- Simular dónde caerían los límites de calificación según ciertos posibles enfoques seleccionados de establecimiento de límites recomendados estadísticamente que fueran factibles, en una serie de asignaturas que capturen toda la gama de contextos que son importantes para el IB
- Revisar la similitud de los resultados simulados en cada enfoque, y los límites recomendados estadísticamente reales y los límites de calificación fijados por la práctica, y extraer conclusiones sobre qué procedimientos podrían ser los más adecuados o inadecuados para los contextos del IB.

Se seleccionaron doce asignaturas para su modelización, que cubrían la amplia gama de contextos de calificación a los que se enfrenta normalmente el IB.

Programa	Asignatura	Contexto de concesión de calificaciones
PD	Matemáticas	Asignaturas estables y grandes
PD	Armenio A: Literatura	Asignaturas muy pequeñas
PD	Sueco A: Literatura	Asignaturas estables y pequeñas
PD	Inglés A: Lengua y Literatura	Asignaturas en crecimiento: crecimiento gradual
PD	Inglés A: Literatura	Asignaturas en descenso
PD	Política Global	Asignaturas en crecimiento: crecimiento importante
PD	Tecnología de la Información en una Sociedad Global (TISG)	Asignaturas en crecimiento: crecimiento rápido
PD	Cine	Cambios de currículo y modelos de evaluación

PD	Ciencias del Deporte, el Ejercicio y la Salud	Nuevas asignaturas: se ha introducido el NS
PAI	Matemáticas	Grupo nuevo en una asignatura ya existente
PD	Teatro	Modelo de “verificación”
PD	Chino B	Distribuciones asimétricas

Se modelizaron cinco enfoques del establecimiento de límites recomendados estadísticamente:

1. **Mantener el estándar anterior:** Los límites del año en curso se han fijado aproximadamente lo más cerca posible de la distribución de las calificaciones del año anterior.
2. **Centros comunes:** En lugar de conservar el resultado de todo el grupo de estudiantes, los grupos de referencia y del año actual se asignan a un subgrupo definido (aquí, un grupo de centros presentes en ambos años). El resultado de solo ese grupo es el que se compara con el grupo de referencia del año en curso (Pinot de Moira, 2019).
3. **Centros comunes estables:** Esta es una variante del método anterior. Aquí el conjunto de centros definidos como comunes se ajusta aún más recurriendo a criterios adicionales. Como tal, el único requisito adicional para este método es el de los criterios definidos por los que los centros se definen como “estables”.
4. **Equiparación de arco de circunferencia:** Utiliza un gráfico de las puntuaciones de la evaluación en curso frente a las de la evaluación de referencia, y dibuja un arco de circunferencia uniendo tres puntos: la puntuación máxima, la puntuación media lograda y la puntuación mínima (Livingston y Kim, 2009).
5. **Resumen instantáneo de logros sin calificaciones (ISAWG, por sus siglas en inglés):** Un enfoque basado en la predicción que emplea los logros simultáneos como indicador externo de las diferencias entre los grupos de estudiantes (Benton, 2017). Dicho de modo resumido, utiliza una amalgama de todas las puntuaciones de los componentes para deducir un indicador general de la capacidad del alumno/a en toda la serie. Esto se hace para los grupos de referencia y del año en curso, y luego se equiparan los dos grupos para establecer una relación entre un año y otro. La medida ISAWG equiparada resultante se puede utilizar luego en lugar de los logros anteriores para predecir los resultados.

Conclusiones

Nuestras conclusiones generales sobre los métodos que son más adecuados para cada contexto se pueden resumir así:

- En asignaturas muy pequeñas de 30-50 estudiantes o menos, el único método viable es la equiparación de arco de circunferencia.
- En asignaturas en crecimiento o en descenso con un crecimiento considerable (con una diferencia anual de alrededor del 25-33 % o más alta), el ISAWG parece el mejor enfoque, aunque los enfoques de centros comunes sean viables si existen datos suficientes para respaldar la subdivisión del alumnado (es decir, dado el tamaño de los grupos del IB, es más fácil que el de centros comunes sea viable).
- En evaluaciones nuevas o que presentan cambios, el factor clave es definir a qué valor de referencia se asigna la asignatura —la metodología concreta de establecimiento de límites recomendados estadísticamente no es tan importante—.

- En otros contextos no mencionados aún, existen diferencias mínimas entre los métodos y es difícil identificar qué es “lo mejor” debido a la falta de una “verdad” objetiva. A menudo todos los métodos se desvían por una distancia y dirección similares de los límites reales fijados. Sin embargo, si aplicamos otro método para explicar las asignaturas en crecimiento o en descenso, tendría sentido utilizar también ese método en otros contextos para reducir la posibilidad de que un cambio imprevisto en la capacidad del grupo tenga un efecto adverso en el estándar.
- A grandes rasgos, la opción entre ISAWG y cualquier otro enfoque no es más que el dilema entre pragmatismo y rigor metodológico, dada la similitud de sus límites recomendados estadísticamente —ISAWG es el único método que permite de manera explícita cambios en la capacidad “en un centro común” dentro de una asignatura, pero es considerablemente más complejo de implementar—. Una posibilidad sería implementar centros comunes como una “victoria rápida” (posiblemente con centros comunes estables cuando esto hace que se incluyan suficientes centros y estudiantes en el modelo), y seguir investigando el ISAWG.

Bibliografía

BENTON, T. "Pooling the totality of our data resources to maintain standards in the face of changing cohorts". Ponencia del 18.º Congreso anual de la Asociación para la Evaluación Educativa de Europa de noviembre de 2017, Praga (República Checa).

ORGANIZACIÓN DEL BACHILLERATO INTERNACIONAL. *Principios y prácticas de evaluación del IB: evaluaciones de calidad en la era digital*. Cardiff (Reino Unido). 2018.

LIVINGSTON, S. A.; KIM, S. "The Circle-Arc Method for Equating in Small Samples". En *Journal of Educational Measurement*. 2009, vol. 46, n.º 3. Pp. 330-343.

PINOT DE MOIRA, A. "Common Centres: In the context of maintenance of standards for the GCSE". Informe inédito para WJEC y CCEA, 2019.



Unit 109 Albert Mill
10 Hulme Hall Road
Castlefield
Manchester (Reino Unido)
M15 4LY

www.alphaplus.co.uk

