Aprender

SERIE DE DOCUMENTOS TÉCNICOS/ 5

# APRENDER 2016
# BOOKMARK
## ESTABLECIMIENTO DE PUNTOS DE CORTE

Secretaría de Evaluación Educativa

Ministerio de Educación y Deportes
Presidencia de la Nación

SERIE DE DOCUMENTOS TÉCNICOS/ 5

# APRENDER 2016
# BOOKMARK
# PUNTOS DE CORTE

# AUTORIDADES

**Presidente**
Ing. Mauricio Macri

**Ministro de Educación y Deportes**
Lic. Esteban Bullrich

**Jefe de Gabinete del Ministerio de Educación y Deportes**
Dr. Diego Marias

**Secretaria de Evaluación Educativa**
Prof. Elena Duro

**Secretario de Gestión Educativa**
Lic. Maximiliano Gullmanelli

**Secretario de Políticas Universitarias**
Dr. Albor Cantard

**Secretaria de Innovación y Calidad**
Sra. María de las Mercedes Miguel

**Secretario de Deportes, Educación Física y Recreación**
Sr. Carlos Mac Allister

## EQUIPO RESPONSABLE

**Coordinación**
Prof. Elena Duro

**Redacción del informe**
Dra. María Aranguren
Lic. Augusto Hoszowski

# ÍNDICE


**PRÓLOGO** 8

**INTRODUCCIÓN** 10

**1. ESTABLECIMIENTO DE ESTÁNDARES** 13
1.1 Sistema de estándares: ¿qué es? ¿para qué sirve? 13
1.2 Evaluación Normativa y Evaluación Criterial 14

**2. ETAPAS Y TAREAS DEL PROCESO DE ESTABLECIMIENTO DE ESTÁNDARES** 17
2.1 Etiquetas y descriptores de los niveles de desempeño 18

**3. VALIDEZ Y CONFIABILIDAD DE LOS MÉTODOS DE ESTABLECIMIENTO DE ESTÁNDARES** 21

**4. APROXIMACIONES METODOLÓGICAS Y FUNDAMENTOS** 25
Una breve introducción a la teoría clásica de los tests y la Teoría de Respuesta al Ítem.

**5. ANTECEDENTES DE LOS MÉTODOS PARA EL ESTABLECIMIENTO DE ESTÁNDARES** 28
5.1 Clasificación de los métodos de establecimiento de estándares 29
5.1.2 Método Angoff y sus variaciones, el método BoW y el método de juicio analítico 30
5.1.3 Comparación del método Angoff, el método BoW, el método de juicio analítico y el método Bookmark siguiendo los criterios enunciados por Berk (1986) 33
5.2 El sustentante limítrofe 34

**6. MÉTODO BOOKMARK** 35
6.1 Descripción del método 35
6.2 Antecedentes de sus aplicaciones 37
6.3 Materiales y procedimientos del método Bookmark 39
6.3.1 Cuadernilo Bookmark y cuadernillo complementario 39
6.3.2 Probabilidad de respuesta 40
6.3.3 Estimación de los puntos de corte 41
6.3.4 Capacitación de los panelistas 42
6.3.5 Composición del panel de participantes 44
6.3.6 Rondas de trabajo 44
6.4 Definición de los puntajes de corte 49
6.5 Finalización del procedimiento Bookmark: redacción de los descriptores de los niveles de desempeño 50

**7. MÉTODO BOOKMARK EN APRENDER 2016** 53

**8. CONCLUSIONES Y RECOMENDACIONES FINALES** 55

**9. REFERENCIAS** 56

**DOCENTES QUE PARTICIPARON DEL TALLER BOOKMARK** 63

# PRÓLOGO

La educación tiene un rol central en el desarrollo social y económico del país. Es nuestra obligación como funcionarios del Estado mejorar de forma continua los niveles de equidad y la calidad del sistema educativo. Para lograr estos objetivos, es fundamental contar con información confiable que guíe la toma de mejores decisiones en los distintos niveles de gestión.

Aprender 2016 generó un conjunto robusto de datos que permite tener una mirada más aguda sobre las fortalezas y desafíos que tenemos por delante. La evaluación por sí sola no corrige los problemas, pero nos brinda valiosa información para profundizar nuestra mirada y continuar encarando un cambio a través de acciones concretas para mejorar la calidad educativa en la Argentina.

Estamos llevando adelante una tarea que nos encomendó el Presidente: generar igualdad de oportunidades para aprender independientemente del lugar donde hayamos nacido.

La evaluación nacional de aprendizajes nos mostró grandes retos en distintas áreas del conocimiento. Al mismo tiempo, nos permitió visualizar un conjunto de escuelas con buenos niveles de desempeño pese a encontrarse en las áreas más vulnerables de nuestra sociedad. Es útil detectar qué cosas funcionan bien para aprovechar todas las experiencias educativas positivas que suceden en nuestro país.

Para concretar semejante tarea es importante destacar que fue invalorable el profundo compromiso demostrado por todos los ministros del país, los equipos de cada provincia, los directores que actuaron como veedores, los docentes aplicadores, los estudiantes y sus familias.

A partir de la información relevada en los resultados de Aprender, estamos trabajando con todas las provincias en políticas de apoyo a las escuelas con niveles de desempeño más bajos con la mirada puesta en la mejora. Es por eso que presentamos el Plan Maestr@, con metas claras para pensar la educación en la actualidad y en el largo plazo.

Con la transparencia que nos caracteriza, compartimos los resultados con la sociedad. Creemos que estos informes alentarán un debate permanente que alimente la revolución educativa que llevamos adelante. Estamos convencidos de que la educación nos une y es una herramienta que contribuye a construir el futuro que todos anhelamos para la Argentina.

La educación nos une.

Lic. Esteban Bullrich
Ministro de Educación y Deportes

# INTRODUCCIÓN

La educación es un derecho y constituye una política central que habilita a ciudadanías plenas, abre puertas al ejercicio de otros derechos y posibilita el desarrollo y crecimiento de la sociedad. La generación de evidencia sólida y confiable sobre el estado de situación de la educación en nuestro país es un factor necesario —aunque no suficiente— para reorientar políticas y prácticas de cara a procesos de mejora educativa continua. El objetivo es realizar un aporte para garantizar una educación de calidad para todos.

La implementación de Aprender 2016 en más de 28 mil escuelas de todo el país, la participación de más de 950 mil estudiantes en esta evaluación, el alto compromiso que han manifestado las autoridades educativas jurisdiccionales, docentes, estudiantes y familias permite hoy devolver información valiosa, en tiempo y forma, a cada una de las escuelas participantes, a los decisores y a la sociedad. Esta información se constituye en un insumo fundamental para conocer y analizar ciertos aprendizajes prioritarios, notas sobre el contexto en el que los estudiantes aprenden y algunos factores que pueden incidir en sus procesos de aprendizaje.

La evaluación educativa es un recorte de una realidad mucho más compleja y por sí sola no mejora los aprendizajes, pero la información que genera, con apoyo y asesoramiento para su uso, se constituye en una herramienta potente a la hora de reflexionar en torno a las prácticas pedagógicas e institucionales, incita a elaborar nuevas preguntas, así como a redireccionar o elaborar prácticas y políticas con el objetivo de mejorar la educación.

La información en torno a Aprender 2016 pone en evidencia importantes desafíos que como sistema educativo hay que enfrentar y superar. Los datos nos muestran que aún es insuficiente la cantidad de estudiantes cuyos resultados se ubican en los niveles de desempeño satisfactorio y avanzado en todas las áreas. Por otro lado, prevalecen altos niveles de inequidad y desigualdad educativa a nivel nacional y jurisdiccional.

Al mismo tiempo, Aprender 2016 pone en evidencia la importancia de la educación como herramienta para alcanzar mayor justicia social. Los resultados muestran cómo todos los estudiantes pueden lograr mejores desempeños, incluso en los contextos más desaventajados, evidenciando el valor agregado que la escuela aporta a los niños y adolescentes de nuestro sistema educativo. Ese aporte es el que se pretende destacar y potenciar a partir de la devolución de los resultados de la evaluación.

El presente documento es parte de un conjunto de informes que incluye: reportes de resultados a nivel nacional y jurisdiccional, informes técnicos y temáticos. Por primera vez, las 24 jurisdicciones del país recibieron, luego de un operativo de esta magnitud, sus informes de resultado a cinco meses de la implementación del dispositivo de evaluación nacional.

Por otro lado, la totalidad de escuelas participantes de Aprender recibe un informe de resultados que hace foco en el potencial de mejora de cada una.

La finalidad de la información de Aprender es colocar la evidencia al servicio de las escuelas, las autoridades, la academia, la comunidad educativa y aportar al debate educativo con miras a traccionar mejoras en la calidad y equidad de la educación argentina.

Prof. Elena Duro
Secretaria de Evaluación Educativa

# 1. ESTABLECIMIENTO DE ESTÁNDARES

## 1.1. SISTEMA DE ESTÁNDARES: ¿QUÉ ES? ¿PARA QUÉ SIRVE?



El desarrollo de un sistema de estándares responde a la necesidad de evaluar la calidad educativa y los procesos de aprendizaje de los estudiantes. Establecer un sistema de estándares garantiza un cierto nivel educativo y la adquisición de determinadas habilidades, capacidades y contenidos homogéneos para todos los estudiantes certificados[1] (Brown, 2001; Cizek, 2006). En este sentido, el desarrollo de los estándares pretende precisar cuáles son aquellas habilidades, capacidades y contenidos que definen cada uno de los niveles de estándares (e.g. nivel bajo, nivel medio, nivel alto), permitiendo evaluar el desempeño de los estudiantes y, en algunas ocasiones, la labor educativa.

El hecho de pensar y trabajar sobre cuáles son esos estándares requeridos para obtener una certificación o título, unifica los criterios de evaluación a nivel provincial y nacional, y permite exponer claramente los objetivos prioritarios de cada nivel educativo a los maestros, directivos y agentes educativos y esto redunda en una mejora de la calidad educativa. El estado (o la organización a cargo) puede, además, difundir cuáles son los criterios propuestos de manera que la sociedad también forme parte y conozca cuáles son las metas y características que debe presentar un estudiante que certifica un determinado nivel educativo (Herrera Ortiz, Benavides Posadas & Monroy Cazorla, 2009).

Existen también otras ventajas por las que resulta conveniente establecer un sistema de estándares. Según Linn (1994), permite motivar a los estudiantes y docentes presentándoles un desafío. En efecto, habiendo objetivos claros de aprendizaje pueden identificarse también las mejoras en los niveles de desempeño. En segundo lugar, muestra claramente las metas y expectativas académicas. En tercer lugar, garantiza que todos los estudiantes que tengan una certificación de determinado nivel presenten ciertos contenidos, habilidades y capacidades mínimas.

De acuerdo a lo anterior, se observa que el establecimiento de estándares es un proceso complejo y no unilateral, dado que implica la participación de distintos actores sociales y los resultados obtenidos influirán en diferentes sectores de la comunidad. Este proceso, entonces, requerirá no sólo de la pericia psicométrica con la que sea abordado sino también de la capacidad para integrar al saber psicométrico los elementos políticos y culturales sobre los que tendrán consecuencias esos estándares (Cizek, 2001).

Una evaluación basada en un sistema de estándares delimita aquellos contenidos, habilidades y capacidades que un estudiante debe dominar. En este sentido, las evaluaciones criteriales buscan conocer cuáles son los elementos mínimos de competencia, y no determinar qué es lo que puede hacer un individuo en comparación a otro (Herrera Ortiz et al., 2009). A nivel internacional, existen diferentes pruebas de criterio bastante reconocidas,

1 Alumno certificado refiere a cualquier estudiante que reciba un título al finalizar un grado académico.

tales como las pruebas desarrolladas por el Programa para la Evaluación Internacional de Alumnos (Programme for International Student Assessment, PISA por sus siglas en inglés) y las pruebas realizadas por el Laboratorio Latinoamericano de Evaluación de la Calidad de la Educación (LLECE), PERCE (Primer Estudio Regional Comparativo y Explicativo, 1997), SERCE (Segundo Estudio Regional Comparativo y Explicativo 2006) y TERCE (Tercer Estudio Regional Comparativo y Explicativo 2013) coordinadas por la oficina regional de educación de la UNESCO para América Latina y el Caribe, entre otras.



El desarrollo de pruebas criteriales en la región, y en la Argentina en particular, es relativamente reciente y el trabajo en relación a la documentación y registro que evidencian la validez de la aplicación de estas pruebas aún es incipiente. En nuestro país el primer Operativo Nacional de Evaluación (ONE) fue realizado en 1993. Otros países como Estados Unidos llevan ya algunos años y experiencia en este campo de la evaluación.

Con el objetivo de esclarecer en qué consisten las evaluaciones criteriales y su relación con el establecimiento de estándares se presenta a continuación información relativa a los diferentes aspectos y elementos que constituyen este proceso.

## 1.2. EVALUACIÓN NORMATIVA Y EVALUACIÓN CRITERIAL

La evaluación psicológica y la evaluación educativa se asientan sobre la idea de que es posible medir cualidades y constructos inobservables. La psicometría es la disciplina que se encarga de diseñar, evaluar y adaptar instrumentos destinados a este propósito.

En los últimos años se ha desarrollado un tipo particular de pruebas: las llamadas pruebas criteriales o tests relacionados con los criterios. Se suelen diferenciar éstos de los tests normativos. En una evaluación normativa se busca identificar cuál es la posición de un estudiante en relación a los demás estudiantes evaluados, mientras que en una evaluación criterial se busca conocer el dominio que tiene un estudiante de ciertos criterios preestablecidos. Normalmente, los criterios de evaluación de ese desempeño son fijados por diferentes jueces que se encargan de esta tarea utilizando alguna de las metodologías y procedimientos para el establecimiento de estándares.

La evaluación criterial pretende corregir las inadecuaciones de los sistemas normativos utilizados en la construcción de tests psicométricos para la elaboración de pruebas estandarizadas (Jornet Meliá & González Such, 2009). El fundamento principal que explica este cambio es la necesidad de que las pruebas puedan ser interpretadas en relación a determinados criterios absolutos de calidad de aprendizaje.

Las primeras definiciones de la evaluación criterial fueron señaladas por Glaser (1963). Una prueba criterial sirve para obtener mediciones directamente interpretables en términos de realizaciones estándar concretas, es decir, lo que el estudiante evaluado puede o no realizar. Otros autores señalan que estas pruebas son utilizadas para determinar la posición de un individuo respecto a un dominio bien definido (Popham, 1978). En este sentido, Popham y Husek (1969), aclaran que el término criterio refiere a un área precisada de modo adecuado y no únicamente al hecho de establecer un estándar de logro o puntos de corte que delimiten dichos estándares.

Las evaluaciones criteriales son utilizadas para evaluar el proceso de aprendizaje y el desempeño de los estudiantes, evaluar programas educativos, identificar dificultades y logros en el

aprendizaje de los estudiantes, y también para evaluar las competencias necesarias a fin de otorgar un título o una licencia (Hambleton & Swaminathan, 1978). En particular, a los organismos que administran estas pruebas les interesa conocer el grado con que el evaluado cumple con los objetivos educativos necesarios, en relación tanto al conocimiento de contenido como con las destrezas y capacidades -conocimientos procedimentales-, que se supone han de conseguir mediante la enseñanza.

Las pruebas criteriales, en efecto, se constituyen de reactivos que responden a los objetivos del aprendizaje. Los ítems que conforman la prueba son un conjunto representativo de un dominio claramente definido (Hambleton & Swaminathan, 1978). De acuerdo con Leyva Barajas (2011) algunas de las características de las pruebas de criterio son: (a) presentan una definición clara y precisa del dominio a evaluar; (b) permiten conocer la ubicación de un sujeto en un continuo representativo del dominio evaluado y; (c) permiten obtener una interpretación directa de la puntuación obtenida: el desempeño que obtiene el estudiante indica su grado de competencia.



En la elaboración de una prueba criterial es posible identificar cuatro operaciones fundamentales: (a) la especificación del dominio a evaluar; (b) el diseño y análisis de los reactivos a incluir; (c) el establecimiento de los puntos de corte y niveles de desempeño (también llamados: niveles de logro) y; (d) la documentación y análisis de las evidencias de confiabilidad y validez de las pruebas (Leyva Barajas, 2011). Dichos procesos son iterativos, es decir, que se repiten y retroalimentan entre sí, hasta obtener resultados satisfactorios. La especificación del dominio refiere a la estructuración y delimitación del área que será evaluado y se relaciona de manera inmediata a la validez de contenido y de constructo que presente la prueba. En la medida que los reactivos de una prueba muestren de manera representativa el universo que desea evaluar, la validez de dicha prueba estará garantizada. A mayor ambigüedad en el contenido de la prueba, mayor será la generalidad de la prueba y menor su validez.

Una vez definidos los ítems que se incluirán en la prueba[2], el segundo paso será establecer los puntos de corte. En este sentido, este manual tiene por interés principal profundizar en los aspectos referidos a esta etapa del proceso. El establecimiento de estándares permitirá el monitoreo y la toma de decisiones de las intervenciones y políticas educativas que se realicen. De acuerdo con Leyva Barajas (2011) "un estándar es un punto en la escala de puntuaciones de una prueba que sirve para clasificar a quienes fueron examinados en categorías que reflejan diferentes niveles de logro referidos a los contenidos y competencias evaluadas" (p. 138).

Concretamente, el establecimiento de estándares consiste en aplicar un sistema objetivo y racional de reglas o procedimientos que permitirán agrupar un número de reactivos capaces de diferenciar entre dos o más niveles de logros (Cizek, 1993). En líneas generales, se podría decir que el establecimiento de estándares tiene como objetivo delimitar y describir distintos niveles de logro o de desempeño, en los cuales pueden ser clasificados los individuos en función de su desempeño.

2 En la mayoría de los procedimientos utilizados para el establecimiento de estándares, antes de establecer los puntos de corte, será necesario, además de definir el conjunto de ítems que integran la prueba, administrarla a los participantes. Esto se debe a que muchos de los métodos que permiten definir los puntos de corte, utilizan el desempeño de los participantes como insumo o elemento de trabajo a partir del cual se establecen los puntos de corte. En el método Bookmark, por ejemplo, será necesario la aplicación de la prueba previamente al establecimiento de los puntos de corte con el objetivo de conformar los Cuadernillos de Items Ordenados (ver en p. 31 Apartado 6.1 Descripción del método).

Los niveles de logro reflejan el nivel de dominio que presentan los estudiantes de los contenidos y habilidades evaluados. Algunos de los más comúnmente utilizados son, por ejemplo: Insatisfactorio, Satisfactorio, Muy Satisfactorio; Insuficiente, Básico, Avanzado; Bajo, Medio, Alto. Los contenidos evaluados pueden ser diversos y cada uno de los niveles de logro incluye a su vez diferentes descriptores que reflejan las habilidades, capacidades y contenidos específicos requeridos para cada dominio o área a la que se esté refiriendo. Normalmente, las categorías o niveles de logro son más informativos que las puntuaciones de las pruebas en sí mismas dado que permiten mostrar el grado de avance de un estudiante de una manera global: caracterizándolo según determinadas aptitudes y logros y diferenciándolo además de las otras categorías existentes (Hambleton, 2001).

Los niveles de logro hacen referencia entonces tanto a las categorías o etiquetas como a las descripciones y definiciones de los contenidos, habilidades y capacidades que se espera encontrar en los examinados de ese nivel. Uno de los objetivos centrales de cualquier evaluación educativa es poder situar a un estudiante en una categoría determinada. De ahí que otro de los objetivos centrales sea determinar cuáles serán esas categorías o niveles de desempeño. Además, los niveles de logro representan la referencia principal sobre la cual los panelistas establecerán los puntos de corte (Herrera Ortiz et al., 2009). El establecimiento de uno o más puntos de corte permitirá determinar aquel o aquellos valores numéricos de la prueba que separen a los examinados en diferentes categorías de acuerdo a su desempeño (Hambleton, 2001). En consonancia con ello, los puntajes de corte indican junto con los descriptores de cada nivel cómo han de ser interpretadas las puntuaciones.

El establecimiento de los puntajes de corte es una parte relevante e influyente en la validez que tengan las interpretaciones de la prueba (Cizek, Bunch & Koons, 2004), de ahí la importancia que tiene una cuidadosa evaluación al momento de elegir cuál será el método seleccionado para el establecimiento de los puntajes de corte. En la actualidad es posible identificar una gran variedad de métodos para el establecimiento de estándares y puntos de corte. La elección de un método dependerá del tiempo y recursos disponibles, del tipo de ítems que se incluyan en la prueba (ítems de respuesta abierta o de respuesta cerrada o ambos) y de la capacidad de los jueces que intervendrán en el establecimiento de los puntos de corte.

# 2. ETAPAS Y TAREAS DEL PROCESO DE ESTABLECIMIENTO DE ESTÁNDARES



El establecimiento de estándares usualmente comprende cuatro tareas básicas: (1) selección de la cantidad de niveles de desempeño que se desean establecer, (2) elección de los nombres o etiquetas de cada nivel, (3) redacción de los descriptores de cada nivel y (4) establecimiento de los puntajes de corte correspondientes.

En la actualidad la mayoría de los procedimientos utilizados para el establecimiento de estándares busca determinar tres o más niveles de desempeño. De hecho, en las sesiones del acto del Congreso de los Estados Unidos *No child left behind* durante el año 2001, se recomendó enfáticamente que cualquier evaluación educativa contara al menos con tres niveles de desempeño. El establecimiento de más de cuatro niveles, sin embargo, también puede provocar algunas dificultades dado que la capacidad para diferenciar entre una categoría y otra se torna más difícil y puede resultar sutil (Perie, 2008). De acuerdo con estas observaciones, sería conveniente establecer tres o cuatro niveles de desempeño pero no más.

Las etiquetas de los niveles se identifican con los términos con los que se reconoce cada una de las categorías de logro (e.g. Desempeño Bajo, Desempeño Medio, Desempeño Alto). En algunos casos se sugiere que la elaboración de las etiquetas y descriptores de cada nivel esté a cargo de los actores políticos y gubernamentales, y se realice teniendo en consideración la agenda y la planificación de la política educativa (Perie, 2008). En otros casos se recomienda que las etiquetas sean desarrolladas por el sector político y gubernamental, y los descriptores diseñados por los jueces que participen en los procedimientos de establecimiento de estándares (Jornet Meliá & Backhoff, 2008). Una tercera alternativa sugiere que el equipo a cargo de la elaboración de las etiquetas desarrolle algunas definiciones teóricas –o también llamadas, descripciones guía– de lo que se espera encontrar en cada nivel de manera genérica. Estas definiciones no han de estar ligadas aún al contenido de la prueba, sino que son principios más generales que se acercan a los lineamientos de la política educativa que guía la evaluación (Herrera Ortiz et al., 2009; Perie, 2008). La idea de esta tercer alternativa es que, en una segunda instancia, estas definiciones genéricas o descripciones guías sean complementadas con los contenidos y habilidades específicas reflejadas en los reactivos de la prueba. En cualquiera de los casos, lo fundamental será que la definición de los niveles –que estará dada por la redacción de sus descriptores– atienda a los propósitos que se tienen al realizar esa evaluación y clasificación de los estudiantes.

La cuarta tarea hace referencia al establecimiento de los puntos de corte. Estrictamente hablando, no existen métodos para el establecimiento de estándares sino que lo que existen son métodos para el establecimiento de los puntos de corte. Estos últimos son los valores numéricos que operacionalizan los niveles de desempeño establecidos. Existen diversos métodos orientados a esta tarea. Cada método posee además un sistema preestablecido y racional de reglas destinadas a identificar cuál es el valor numérico adecuado que separa un nivel de desempeño de otro. Algunos de los más conocidos son el método de Angoff (1971) y el método Bookmark (Lewis, Mitzel, & Green, 1996; Lewis, Mitzel, Green, & Patz, 1999).

Independientemente del método que se utilice para el establecimiento de puntos de corte, todo procedimiento de establecimiento de estándares involucrará:

(a) selección del método,
(b) selección de jueces,
(c) definición del individuo limítrofe,
(d) capacitación de los jueces en el método seleccionado,
(e) obtención de los puntos de corte establecidos por los grupos de jueces,
(f) retroalimentación de puntos de corte preliminares (Herrera Ortiz et al., 2009).



## 2.1 ETIQUETAS Y DESCRIPTORES DE LOS NIVELES DE DESEMPEÑO

La selección de los nombres o etiquetas de cada nivel implica la elección de términos que refieren al nivel de dominio que tiene un estudiante sobre un área de competencia. El propósito de asignar un nombre es proporcionar información y un marco de referencia común a los jueces que redactarán las descripciones (Herrera Ortiz et al., 2009). Además, las etiquetas permiten sintetizar una serie de ideas en una sola palabra y simplifica la comunicación. Existen varias nominaciones para las categorías de desempeño que varían de acuerdo a los usos y costumbres de cada contexto cultural (e.g.: Básico, Competente, Avanzado; Elemental, Satisfactoria, Sobresaliente; Limitado, Básico, Acelerado, Avanzado; Lejos de Suficiente, Por debajo de Suficiente, Suficiente, Competente, Avanzado; No alcanza los objetivos, Alcanza los objetivos, Alcanza los objetivos satisfactoriamente). Acorde a ello, las etiquetas son seleccionadas de modo que representen mejor las categorías que se postulan.

Algunos criterios para la selección de etiquetas son: (a) simplicidad; (b) capacidad discriminativa entre las categorías o niveles; (c) neutralidad en términos de valoración correcta/incorrecta, superior/inferior; (d) claridad y precisión; (e) flexibilidad para ser revisadas (Jornet Meliá & Backhoff, 2006). Las etiquetas deben ser simples y fácilmente comprensibles. En este sentido, se espera que sean claras y precisas, que eviten la ambigüedad, los términos técnicos y/o confusos. Por otra parte, deben poder diferenciar claramente tipos de desempeños diferentes, evitando utilizar términos que puedan ser interpretados de manera negativa (e.g. "inferior", "reprobado", "desaprobado", entre otros). También se recomienda evitar términos que impliquen movilidad tales como: "cercano al promedio" o "en progreso", entre otros (Beck, 2003). De ser necesario, las etiquetas pueden ser modificadas a partir de la aplicación de la prueba y de sus resultados.

Una vez que se establece la cantidad de niveles y las etiquetas, pueden ya redactarse los descriptores de cada nivel. Éstos hacen referencia a los conocimientos, habilidades y capacidades que los estudiantes deben dominar según el nivel de desempeño en el que se posicionan. Es decir, en pocas palabras establecen el significado y alcance de los puntajes de corte. De esta manera, ayudan a los maestros, padres y actores políticos a interpretar qué es lo que los estudiantes de cada nivel conocen y pueden hacer, y potencialmente, qué no conocen y no pueden hacer.

Cualquier procedimiento para el establecimiento de corte debe analizar cuidadosamente cuáles son los descriptores de cada nivel. Un valor numérico que no referencia a un contenido específico es parcial e insuficiente. Los valores numéricos de los puntajes de corte deben reflejar diferencias en los contenidos y conocimientos dominados por los distintos grupos. En cierto sentido, los descriptores de cada nivel de desempeño son el fundamento

de los criterios que serán utilizados en el establecimiento de estándares dado que proveen una explicación pormenorizada de los contenidos, habilidades y capacidades que se esperan encontrar en cada nivel (Perie, 2008).

Idealmente una evaluación de desempeño debería tener definidos los niveles de desempeño y los objetivos de utilizar esos niveles. Incluso, esos mismos objetivos y propósitos deberían ser tenidos en cuenta para el diseño de los ítems de la prueba de modo que ésta responda a lo que las instituciones políticas y/o gubernamentales o, en algunos casos, organismos sociales desean conocer (Perie, 2008). Comúnmente, sin embargo, los descriptores de los niveles de desempeño son elaborados posteriormente al desarrollo de los ítems y, de manera previa o inmediatamente posterior, al establecimiento de los puntajes de corte.



En este sentido, es usual que se establezca la cantidad de niveles de desempeño, se seleccionen las etiquetas, se diseñe un pool de ítems que responda a cada área y tenga en cuenta esas etiquetas; luego, se establezcan los puntos de corte y, como parte del proceso de determinación de los puntos de corte, se redacten los descriptores. Los jueces que participan en las tareas correspondientes al establecimiento de puntos de corte deben entonces considerar cuáles son esos descriptores de niveles de desempeño teniendo en cuenta, por un lado, lo que a nivel político y gubernamental se desea evaluar y, por otro, el alcance de los contenidos incluidos en los ítems redactados.

En este sentido, Perie (2008) señala que sería más fácil desarrollar las etiquetas y descriptores de los niveles de desempeño con anterioridad, de manera que los diseñadores de ítems –especialistas en contenido y equipo pedagógico a cargo–, conozcan cuáles son los niveles de desempeño y a qué se dirigen. De manera contraria, señala la autora que los niveles de desempeño pensados o pretendidos tienen poca influencia sobre el diseño de los ítems y un impacto muy alto sobre la ubicación de los puntajes de corte.

A nivel operativo, cada país y estado utiliza diferentes métodos para desarrollar los descriptores de los niveles y los mismos, tal como se mencionó, pueden ser elaborados en distintos momentos del establecimiento de estándares. La situación más común implica que los descriptores no sean desarrollados hasta el establecimiento de los puntos de corte.

En algunos casos, la redacción de los descriptores se realiza antes del establecimiento de los puntos de corte y puede estar a cargo de los mismos jueces seleccionados para el establecimiento de puntos de corte (Cizek et al., 2004). En otros casos, la elaboración de los descriptores puede estar a cargo de un comité diferente de expertos pedagógicos y conocedores de los contenidos evaluados (Jornet Meliá & Backhoff, 2006, 2008), o incluso de un grupo perteneciente al sector político/gubernamental. Una tercera posibilidad recomienda la elaboración de una descripción guía a cargo de quienes proponen las mismas etiquetas y una elaboración posterior de esos descriptores guías más completa y basada en los contenidos específicos de cada examen a cargo de los jueces seleccionados para implementar el método de establecimiento de puntos de corte (Herrera Ortiz et al., 2009). Esta tercera posibilidad admite que los descriptores sean redactados antes o después del establecimiento de los puntos de corte.

Típicamente, los procedimientos basados en la Teoría de Respuesta al Ítem (TRI) -que utilizan mapas de reactivos-, se inclinan a desarrollar los descriptores de los niveles de desempeño una vez establecidos los puntajes de corte (Perie, 2008). De esta forma, el contenido de los descriptores está en total consonancia con los ítems que constituyen cada nivel. La

desventaja de esta alternativa es que los jueces no tienen una delimitación clara de qué se espera encontrar en cada nivel durante el proceso de establecimiento de los puntos de corte. Esto, a su vez, puede influir directamente en los resultados ya que los panelistas no cuentan con descripciones objetivas que expliciten y acuerden criterios comunes a cada nivel. Al respecto, Cooper-Loomis y Bourque (2001) recomiendan los siguientes pasos: (1) especificar la cantidad y nombres de los niveles de desempeño; (2) esquematizar los descriptores de cada nivel o armar un borrador; (3) confeccionar los descriptores finales para cada nivel una vez establecidos los puntos de corte, alineando los descriptores borradores en función de los indicadores que han sido incluidos bajo cada etiqueta.

Cualquiera sea la alternativa elegida, los encargados de esta tarea deberán estar familiarizados con la amplitud de la especificidad del contenido debido a que dicha amplitud deberá estar reflejada en los descriptores que se redacten.

# 3. VALIDEZ Y CONFIABILIDAD DE LOS MÉTODOS DE ESTABLECIMIENTO DE ESTÁNDARES



Los aspectos psicométricos del sistema de estándares garantizan que las decisiones y clasificaciones resultantes del proceso de evaluación sean más justas y acertadas. Para ello es necesario que los procedimientos utilizados en el establecimiento de estándares evidencien ser válidos y confiables y se asienten sobre criterios y reglas objetivas, eliminado o disminuyendo al mínimo los elementos subjetivos que puedan filtrarse.

Comúnmente se dice que un instrumento es confiable cuando al repetirse su aplicación se obtienen medidas similares o valores cercanos; y se dice que es válido cuando mide específicamente aquello que se propone medir y no otra cosa (Cortada de Kohan, Macbeth & López Alonso, 2008). Es decir, la confiabilidad de una técnica tiene que ver con cómo mide un instrumento un determinado constructo (qué tan bien lo hace) y la validez se refiere a lo que mide y orienta sobre qué se puede inferir a partir de los resultados obtenidos.

Existen distintas maneras de evaluar la confiabilidad y validez de los instrumentos y procedimientos de evaluación. Los procedimientos seleccionados dependerán de las características de los constructos que se estén evaluando y de las pruebas que hayan sido diseñadas para este fin. En el ámbito de la evaluación educativa, la APA (American Psychological Association), la AERA (American Educational Research Association) y el NCME (National Council on Measurement in Education) hacen hincapié en la importancia de incluir programas de investigación que permitan analizar y monitorear la validez de los trabajos realizados (Cizek et al., 2004). Tal como se señala en los Estándares sobre el uso de Tests Psicológicos y Educativos (Standards for Educational and Psychological Testing. APA/ AERA/ NCME, 1999), la validez de las inferencias que se realicen acerca de las puntuaciones obtenidas dependerá de cuán ajustada y válida sea la diferencia entre un adecuado y un inadecuado desempeño. Es decir, la validez de las inferencias dependerá de la validez de los procedimientos utilizados, y la validez de éstos estará dada por los diferentes tipos de evidencia que se recolecten en las etapas que componen el establecimiento de los estándares. Por su parte, la confiabilidad de un procedimiento de medición tendrá que ver con el grado de consistencia entre los resultados obtenidos en distintas mediciones.

En otras palabras, la confiabilidad se referirá al grado en que esos resultados o esas medidas están libres de error (APA/ AERA/ NCME, 1999). El error de medida o el error en los procedimientos de establecimientos de estándares limita la generalización de los resultados.[3]

El establecimiento de los puntajes de corte implica una serie de aspectos a tener en cuenta. Uno de los principales hace referencia a la participación de un grupo de panelistas o

3 En este apartado se hará mención específicamente a las evidencias de validez y confiabilidad de los procedimientos para el establecimiento de estándares. En este sentido, si bien los procedimientos de establecimientos de estándares se realizan sobre las pruebas criteriales administradas a los estudiantes o al grupo de examinados al que competa, no se hará mención a la necesidad de que esas pruebas administradas también presenten sus respectivas evidencias de validez y fiabilidad (e.g. las pruebas criteriales deben incluir indicadores que sean representativos del dominio y que se adecúen a la definición de dominio dada; los ítems deben estar redactados de manera simple, clara e inteligible, entre otros).

jueces que, luego de una cuidadosa evaluación de los ítems que componen la prueba, recomendará uno o más puntos de corte (dependiendo de cuántas categorías o niveles se requieran establecer) basados en su criterio y experticia. En relación a ello, e independientemente del método que se seleccione, Cizek et al. (2004) señalan algunas cuestiones a tener en cuenta para garantizar la validez y confiabilidad de los resultados obtenidos.

Un primer elemento hace referencia a la representatividad y selección de los jueces. Se recomienda que los procedimientos de contacto y selección de los mismos sean cuidadosamente diseñados de modo que la representatividad de la población sea garantizada en dicha selección. La idea que subyace a esta recomendación apunta a obtener resultados confiables, o en otros términos, replicables. De acuerdo con Hambleton (2001), la pregunta que debe hacerse es:

> "Si hubiese otra evaluación de estándares,
> ¿la muestra seleccionada arribaría a los mismos resultados que la primera?".

Los Estándares sobre el uso de Tests Psicológicos y Educativos (APA/ AERA/ NCME, 1999) señalan que tanto la cualidad de los participantes –sus características sociodemográficas– como la cantidad deben orientarse a asegurar que los resultados obtenidos en el establecimiento de estándares por un grupo determinado de panelistas, sea sino el mismo, similar a los que se podrían obtener si los procedimientos fueran aplicados por un segundo grupo muestral seleccionado.

La muestra deberá ser representativa de todos los actores sociales que intervienen en el ámbito educativo y también de las distintas escuelas (privadas, estatales, con subvención del estado, urbanas, rurales, suburbanas). Puede incluir maestros con diferentes grados de experticia y de diferente género, maestros de diferentes grados, maestros especiales, directores de escuela, secretarios del ministerio de educación, entre otros (Hambleton, 2001). La selección de la muestra deberá ser intencional siguiendo un criterio racional.

Un segundo elemento pone en consideración el entrenamiento que reciben los jueces acerca del método y tareas que han de realizar. Por un lado, los panelistas deben comprender claramente cuáles son las actividades y los objetivos de las tareas propuestas. Por otro lado, la capacitación también deberá orientarse a que los participantes adquieran una adecuada conceptualización e identificación del grupo y de sus características donde se aplicarán los estándares establecidos. En especial, cada método para el establecimiento de estándares necesita que los jueces logren identificar lo que se denomina el individuo limítrofe entre una categoría de desempeño y otra. Por ejemplo, en el caso del método Angoff (1971) será necesario que los panelistas asimilen el concepto de sustentante de competencia mínima (*minimally competent examinee*) mientras que para el método Bookmark (Lewis et al., 1996; Lewis et al.,1999) será necesario que los participantes identifiquen al sustentante límite que podría contestar correctamente el ítem con una probabilidad mayor a .67. En relación a esto, algunos estudios han mostrado que cuando los jueces no entienden claramente a qué se hace referencia con el término de individuo limítrofe o sustentante limítrofe, los puntajes de corte recomendados varían en un amplio rango y disminuyen en su confiabilidad (Impara, Giraud, & Plake, 2000).

Para garantizar la validez es indispensable documentar y registrar cada una de las etapas y características referidas a cómo fue llevado a cabo el trabajo realizado. En particular, es importante evaluar el grado de compresión, el nivel de competencia percibido, la au-

toeficacia y seguridad con que los participantes llevaron a cabo la tarea. En este sentido, no sólo se recomienda brindar una adecuada orientación a los participantes (jueces y/o panelistas) acerca del propósito de la tarea y de la agenda de trabajo, sino también evaluar el grado de comprensión una vez brindada la capacitación. Ésta es central para otorgar un marco de referencia al trabajo que los panelistas deben realizar, además, participa a los jueces del proceso y los involucra. Por su parte, la evaluación de la capacitación sirve para garantizar que efectivamente los jueces comprendan el objetivo de su trabajo y se asegure la adecuada ejecución de sus tareas.



También es recomendable valorar la conformidad de los panelistas hacia el final de la agenda de trabajo y una vez finalizado el proceso de establecimiento de puntajes de corte. Esta evaluación es relevante para poder obtener evidencias acerca de la validez de los procedimientos. En este sentido, y dado que el procedimiento aplicado para el establecimiento de estándares es arduo y relativamente largo, Cizek et al. (2004) recomiendan administrar un cuestionario de preguntas cerradas y evitar las preguntas abiertas que exigen mayor elaboración. Básicamente, el objetivo de este cuestionario es conocer si el procedimiento utilizado fue razonable y resultó útil e inteligible para los panelistas. Este cuestionario incluye indicadores que se dirigen a evaluar cuestiones relativas a la utilidad de las etiquetas y de los descriptores propuestos, como de los puntajes de corte finalmente establecidos.

Por último, otro de los aspectos relevantes en el establecimiento de estándares hace referencia a la elección de las etiquetas de los niveles de desempeño y a la confección de los descriptores que componen cada nivel. Tal como se mencionó, en ocasiones las etiquetas y los descriptores pueden ser sugeridos por autoridades gubernamentales de acuerdo a las expectativas y los lineamientos de política educativa que pretenda establecer. Normalmente en estos casos, las etiquetas y los descriptores se encontrarán alineados a los objetivos de calidad educativa y los ítems e indicadores se desarrollarán en función de estos objetivos, etiquetas y descriptores (Perie, 2008). En otros casos, los descriptores de cada nivel pueden ser desarrollados por los panelistas y en relación a ello, también se deberá llevar un adecuado registro y documentación que garantice la validez de los procedimientos utilizados tanto para la determinación de las etiquetas, la elección de la cantidad de los niveles de logros y los descriptores de cada nivel. Independientemente de la persona o el grupo de personas que establezca las etiquetas y descriptores, esto deberá quedar asentado y documentado.

Los Estándares publicados por la APA, AERA y NCME (1999) recomiendan que se tenga en consideración diferentes aspectos ligados a la selección de la muestra de jueces, capacitación y evaluación de las distintas etapas de participación de los panelistas con el propósito de registrar distintos tipos de evidencias de validez y confiabilidad de los procedimientos utilizados para el establecimiento de estándares. En consecuencia, uno de los aspectos clave será la transparencia con que se lleve a cabo el proceso y la documentación que respalde los procedimientos utilizados.

No hay un único procedimiento que sirva para establecer los puntajes de corte de todas las pruebas educativas. De hecho, actualmente existen más de 50 métodos para el establecimiento de estándares (Berk, 1986). Cada uno ha sido propuesto con el objetivo de mejorar los precedentes e incrementar la objetividad (y, por tanto, disminuir la subjetividad o arbitrariedad) de las evaluaciones de los jueces. El método que se adopte dependerá de los objetivos propuestos. En este sentido, será preciso observar: (a) las características de

cada método; (b) los estudios comparativos realizados entre ellos; (c) las características de las pruebas que serán administradas a los estudiantes; (d) la disponibilidad de recursos (de tiempo y económicos), y en función de ello, seleccionar el método adecuado para cada caso (Jornet Meliá & Backhoff, 2008).

Para finalizar, Berk (1986) menciona algunos criterios a tener en cuenta para la selección del método que será utilizado entre los que se señalan unos técnicos y otros que hacen a la factibilidad del método. Los seis criterios que hacen referencia a la pericia técnica del método fueron elaborados por Berk teniendo en cuenta las recomendaciones de la APA, AERA y NCME, como también la opinión de expertos en el área y aspectos jurídicos-legales. Los cuatro criterios restantes están dirigidos a evaluar la factibilidad, es decir, qué tan practicable o viable es el método.

Los criterios de adecuación técnica establecen que el método debe: (a) arribar a una clasificación apropiada de los datos o de la información; (b) ser sensible al desempeño de los examinados; (c) considerar aspectos relacionados con la instrucción y entrenamiento de los jueces; (d) aportar datos y sustento estadístico para la toma de decisiones de los jueces; (e) ser capaz de identificar el "verdadero" estándar y; (f) aportar evidencias de validez para las decisiones que se tomen a partir de los resultados obtenidos. En relación a la factibilidad, Berk (1986) señala que el método debe ser: (g) fácil de implementar; (h) fácil de computar (cálculos estadísticos y/o aritméticos sencillos); (i) fácil de interpretar y; (j) confiable para los no especialistas en el área.

# 4. APROXIMACIONES METODOLÓGICAS Y FUNDAMENTOS. UNA BREVE INTRODUCCIÓN A LA TEORÍA CLÁSICA DE LOS TESTS Y LA TEORÍA DE RESPUESTA AL ÍTEM



Una prueba es un instrumento de medición que presenta características particulares. En este sentido, una de las particularidades de las pruebas psicométricas es que son mediciones indirectas de los constructos o variables a los que se orientan (Cortada de Kohan et al., 2008). Toda prueba psicométrica, además, es diseñada y construida a partir de postulados teóricos que determinan cuál es la mejor manera de medir un constructo. Las teorías de las pruebas permiten estimar las propiedades psicométricas de los tests y establecer cuándo un instrumento es válido y confiable. De este modo, se garantiza que las decisiones tomadas a partir de los resultados de dichos tests sean las adecuadas.

Existen dos grandes teorías referidas a la construcción y diseño de tests: Teoría Clásica de los Tests (TCT) y la Teoría de Respuesta al Ítem (TRI). El enfoque de la TCT es el más tradicional y desde el cual se han construido los instrumentos de medición más utilizados en el ámbito de la Psicología, como por ejemplo, el test gestáltico visomotor de Bender (1938), las diferentes escalas de inteligencia de Wechsler (WISC, 1949), el inventario de depresión de Beck (BDI, Beck & Steer, 1987), entre otros (Muñiz, 2010). La TCT tuvo sus comienzos a principios del siglo XX dando lugar a grandes progresos en el ámbito de la medición, mientras que la TRI tiene sus primeros aportes en la década del '60 gracias a las contribuciones de Rasch (1960) y de Lord y Novick (1968). Ambas teorías tienen como objetivo garantizar que las pruebas utilizadas sean construidas bajo determinados preceptos analíticos y estadísticos que aseguren su confiabilidad y validez.

La intención de una prueba es registrar una serie de respuestas del examinado que reflejen cuánto posee de un determinado atributo (Cortada de Kohan et al., 2008). Dado que las respuestas del examinado son medidas indirectas de los constructos evaluados se dice que siempre hay un margen de error en las mediciones. Éste es el error que los psicometristas buscan conocer en primer lugar, y, en segundo término, reducir. De ahí que los modelos teóricos que subyacen a la construcción de una prueba sean particularmente útiles. Cada modelo teórico propondrá una serie de procesos que permita estimar el grado de precisión (o el grado de error) de las pruebas diseñadas (Muñiz, 2010). En otras palabras, cada modelo diferirá en los procedimientos estadísticos utilizados para separar la puntuación verdadera del error de medida. El mismo surge porque en la evaluación no sólo operan las variables del estudio sino también otras llamadas variables extrañas o contaminantes que pueden influir en los resultados obtenidos.

En breves palabras, se podría señalar que el postulado principal de la TCT es que el puntaje obtenido por un individuo en un test responde a dos componentes: la puntuación verdadera (imponderable) y el error de medida. Un modelo o relación lineal que puede ser representado de la siguiente manera: X = V + e, donde: X = puntuación bruta o empírica de un sujeto (observable); V = valor verdadero de su aptitud (inobservable) y; e = error (Cortada de Kohan et al., 2008). Es decir, la TCT postula que los efectos sistemáticos entre las respuestas de los examinados se deben únicamente a la variación en la habilidad o a diferencias en su variabilidad verdadera. Las demás fuentes posibles de variabilidad (e.g. materiales,

condiciones internas o externas de los examinados) deben, o bien ser mantenidas constantes mediante técnicas de estandarización (e.g. misma consigna, mismo lugar de evaluación, misma fuente y tamaño de letra, idénticas condiciones lumínicas), o bien tienen un efecto que no es sistemático, es decir que se debe al azar. El error no sistemático, el error que se debe al azar, es el que se pretende estimar cuando se evalúa la confiabilidad de un instrumento. Se pretende que este error no sistemático sea el menor posible.

Otra característica de la TCT es que la aptitud se expresa como un puntaje verdadero. Esto es, la habilidad de un individuo se define en función de un test particular. Ahora bien, la limitación de esta manera de comprender la respuesta es que si la prueba es difícil, el examinado parecerá presentar poca aptitud, mientras que si la prueba es fácil, el examinado parecerá presentar mucha aptitud.

La Teoría de Respuesta al Item (TRI) o Teoría del Rasgo Latente es una teoría relativamente reciente y se caracteriza por integrar la evaluación a modelos probabilísticos que permiten conocer la información proporcionada por cada ítem y crear pruebas individualizadas o a medida. De acuerdo con Muñiz (2010), el postulado principal de los modelos TRI es que existe una relación funcional entre los valores de la variable que miden los ítems y la probabilidad de acertar a éstos. Esta relación funcional se llama Curva Característica del Item (CCI). La forma que adopte la CCI dependerá del valor que tomen los tres parámetros que la determinan: (a) el índice de discriminación del ítem; (b) la dificultad del ítem y (c) la probabilidad de contestar correctamente el ítem al azar.

La función matemática que se elija para la CCI condicionará el modelo que se seleccione. Actualmente, existen más de 100 modelos basados en la TRI. La utilización de uno u otro dependerá del tipo de datos manejados (e.g. escalamiento tipo Likert, escalamiento dicotómico, datos multidimensionales) (Muñiz, 2010). Normalmente, las funciones más utilizadas son la logística y la curva normal. Los modelos más utilizados son los logísticos puesto que producen resultados similares pero son más fáciles de usar. De hecho, el primer modelo probabilístico que dio lugar a este tipo de evaluación fue uno logístico, el modelo de Rasch (1960). Éste asume un único parámetro que es el de la dificultad de los ítems (b). Posteriormente, surgieron otros modelos, como el de dos parámetros que tiene en cuenta el índice de discriminación (a), y el modelo de tres parámetros que tiene en cuenta el factor azar en las respuestas a ítems de alternativas múltiples (c).

La diferencia principal que existe entre la TCT y la TRI es que la relación entre el puntaje observado y el rasgo o aptitud de la teoría clásica es de tipo lineal ($X = V + e$), mientras que en los diversos modelos de la TRI, las relaciones son de tipo exponencial, principalmente logístico de uno, dos o tres parámetros; aunque también existen otros modelos como el modelo de Poisson, de la ojiva normal y del error binominal (Cortada de Kohan et al., 2008). La TRI busca dar una fundamentación probabilística al problema de la medición de los constructos inobservables. Estos modelos son funciones matemáticas que relacionan las probabilidades de una respuesta particular a un ítem con la aptitud general del sujeto. En cambio, la TCT sigue siendo utilizada y ha sido de gran de utilidad, pero diversos estudios han señalado dos limitaciones críticas: (a) las mediciones no resultan invariantes respecto del instrumento utilizado y (b) la ausencia de invariancia de las propiedades psicométricas de las pruebas respecto de las muestras utilizadas para estimarlas (Muñiz, 2010).

De acuerdo con Hambleton, Swaminathan y Rogers (1991) en la TCT un ítem es sencillo o complejo de acuerdo a la aptitud de los examinados y ésta depende de que los ítems del

test sean sencillos o complejos, lo que es una paradoja. En otras palabras, para la TCT las características de los ítems son dependientes del grupo.

Los indicadores psicométricos clásicos desarrollados a partir de la TCT no suelen mostrar buenos resultados al ser aplicados a las llamadas pruebas criteriales, por lo que existen nuevas herramientas tecnológicas que se ajustan a las características de estas pruebas permitiendo evaluar su confiabilidad y validez, y determinar los puntos de corte necesarios para establecer el nivel de desempeño que muestra un individuo en un dominio evaluado (Cizek, 2001; Muñiz, 2010).

# 5. ANTECEDENTES DE LOS MÉTODOS PARA EL ESTABLECIMIENTO DE ESTÁNDARES

Existen diferentes métodos para el establecimiento de estándares (Zieky, 1995, 2001) y diferentes maneras de clasificarlos (Berk, 1986; Cizek, et al., 2004; Cizek, 1996; Glass, 1978; Hambleton, Jaeger, Plake & Mills, 2000; Jornet Meliá & Backhoff, 2008; Meskauskas, 1976; Shepard, 1980, 1984). Los mismos han ido variando a través del tiempo y evolucionando de acuerdo a los cambios culturales, sociales y técnicos que inciden finalmente en la posibilidad de instrumentalizar determinados procedimientos y conceptualizar diferentes niveles de logro.

Siguiendo a Jornet Meliá y Backhoff (2008) se pueden distinguir tres momentos en la evolución de los métodos. El primero se caracteriza por el desarrollo de las pruebas criteriales en el ámbito de la evaluación educativa. El diseño de pruebas criteriales en la década del '60 pone de manifiesto la necesidad de contar con métodos para poder valorar de manera adecuada los resultados obtenidos en dichas evaluaciones. En otras palabras, la interpretación de las puntuaciones debe indicar si el estudiante domina o no el conocimiento evaluado. De esa forma, se intentan identificar criterios objetivos que permitan generar adecuadas interpretaciones.

En una segunda etapa, que continúa la anterior, se diseñan algunos métodos más objetivos que permiten obtener algunas interpretaciones de tipo dicotómico (aprueba/no aprueba), y es recién en 1990 cuando se puede reconocer una tercera etapa caracterizada por el interés de desarrollar métodos orientados a determinar múltiples niveles de desempeño y, en consecuencia, varios puntos de corte. En la actualidad, la mayoría de las pruebas criteriales realizadas a nivel nacional e internacional adhieren a este sistema de evaluación politómico en el cual se busca establecer más de dos niveles de desempeño.

Una de las principales razones que ha dado lugar a la pluralidad de métodos para el establecimiento de estándares –y de puntajes de corte– es la arbitrariedad inherente al concepto de calidad (Jornet Meliá & González Such, 2009) y a los métodos basados en los juicios de experto. En las pruebas criteriales –a diferencia de las normativas– no basta saber qué tan bien se desempeñó un sujeto respecto de otro. El objetivo es determinar qué tan bien domina un examinado un conocimiento o una habilidad, y este "qué tan bien" es lo que hace referencia a la calidad y al componente subjetivo de la evaluación. En este sentido, los métodos basados en el juicio de experto han mostrado algunas limitaciones en relación a su capacidad objetiva para determinar los niveles de desempeño. Justamente, la gran parte de los procedimientos adoptados para resguardar la confiabilidad apuntan a poder obtener las mismas interpretaciones y puntuaciones de corte a pesar de la variabilidad que pueda haber en los grupos de panelistas consultados. Lo mismo sucede en relación a la validez. Dado que la calidad de la interpretación dependerá del juicio de experto se debe garantizar que los procedimientos sean lo más rigurosos y exhaustivos posibles, reduciendo de este modo, la posibilidad de error y arbitrariedad en los resultados.

# 5. 1. CLASIFICACIÓN DE LOS MÉTODOS DE ESTABLECIMIENTO DE ESTÁNDARES

Hambleton et al. (2000) ofrecen una clasificación esquemática de los diferentes métodos existentes basándose en la tarea que deben realizar los jueces. En este sentido, los autores señalan que todos los métodos pueden ser considerados en última instancia como métodos de juicio. Lo que varía de uno a otro para esta clasificación es el objeto sobre el cual los panelistas deberán emitir su juicio. Así, de acuerdo al objeto que tengan, las tareas se podrán caracterizar por realizar: (a) juicios basados en una revisión del material y de los indicadores; (b) juicios acerca del trabajo realizado por el examinado; (c) juicios acerca de los perfiles de rendimiento y; (d) juicios acerca de los candidatos.



En el primer grupo, se encuentran los métodos más conocidos y más utilizados como el método de Angoff (1971) y el método Bookmark (Lewis et al., 1996; Lewis et al., 1999). También se menciona el de Ebel (1972), las modificaciones al de Angoff de Hambleton y Plake (1995) y el de Cooper-Loomis y Bourque (2001).

En segundo lugar, entre los métodos en donde los jueces deben realizar un juicio acerca del trabajo de los examinados, Hambleton et al. (2000) ubican al de selección de trabajos (the paper selection) propuesto por la ACT (American College Testing), el del cuerpo de trabajo (the body of work method- BoW) propuesto por Kahl, Crokett, DePascale, y Rindfleish (1994) y Kingston, Kahl, Sweeney y Bay (2001) y el de juicio analítico (analytical judgement) de Plake y Hambleton (1998, 2001). En particular, estos dos últimos son considerados actualmente como métodos más holísticos dado que requieren a los panelistas un juicio acerca de una muestra del trabajo del examinado y no solamente un juicio acerca del desempeño en un ítem específico o la realización de tareas secuenciales que no permiten tener una visión global del fenómeno observado (Cizek et al., 2004).

En tercer lugar, como ejemplos de métodos donde el juicio es acerca de los perfiles de desempeño, se pueden nombrar el del perfil dominante (the dominant profile method) de Jaeger (1995) y el de policy-capturing de Plake, Hambleton y Jaeger (1997). En la cuarta categoría, se puede mencionar el basado en el grupo de referencia (the contrasting group) de Livingston y Zieky (1982).

Por su parte, Jornet Meliá y Backhoff (2008) adoptan otra clasificación según la cual se podrían distinguir tres tipos: (a) de juicio; (b) empíricos y; (c) mixtos. La mayoría de los procedimientos y métodos existentes se ubica dentro de la primera categoría. En éstos, el establecimiento de estándares dependerá de la valoración que realicen los panelistas acerca de los ítems, los sujetos o las tareas. Tal es el caso por ejemplo del método de Angoff (1971), del de Jaeger (1978) o el de Ebel (1962). Entre los empíricos Jornet Meliá y Backhoff (2008) incluyen a aquellos en los que se destaca el uso de criterios estadísticos para promover una mejor decisión. Se puede mencionar como ejemplo los modelos de estado, como lo son los de Emrick y Adams (1969 citados en Jornet Meliá & Backhoff, 2008) y Emrick (1971 citado en Jornet Meliá & Backhoff, 2008), y los continuos basados en la teoría de la decisión que dieron lugar a distintos procedimientos. Por último, los mixtos integran características de los anteriores y es común que, en una primera instancia, se tome en consideración los juicios realizados por los panelistas para luego ajustar estas valoraciones y las puntuaciones de corte finales en función de algunos otros elementos empíricos que tengan en consideración el funcionamiento de las pruebas. Se puede nombrar dentro de este grupo los métodos de compromiso como son los de De Gruijter (1985), el de Hoffstee

(1983) y el de Beuk (1984); y los de correspondencia de ítems como son el Bookmark (Lewis et al., 1996; Lewis et al., 1999).

Como se observa existe en la actualidad una pluralidad de métodos que varían en función de cómo son realizadas las tareas, cuál es elemento preponderante a tener en cuenta en la toma de decisiones acerca de los puntajes de corte que se establecerán, el tiempo requerido para las tareas, los recursos cognitivos y la exigencia de las tareas solicitadas a los jueces, entre otros. La elección de uno u otro método dependerá de:



**a)** la heterogeneidad de los ítems utilizados en la evaluación (e.g. ítems de opción múltiple suelen ser evaluados con el método Angoff, mientras que cuando se utiliza una combinación de ítems de opción múltiple y de preguntas abiertas se suele recomendar el método Bookmark);

**b)** el tiempo disponible para realizar las evaluación;

**c)** experiencia previa con el método (e.g. la utilización previa del método facilita el trabajo de campo posterior y reduce la cantidad de tiempo a las tareas, por lo tanto, a veces es más ventajoso);

**d)** las evidencias de validez del método (e.g. algunos investigadores han cuestionado la validez del método Angoff y del método de grupos contrastados).

Principalmente, los métodos más utilizados hoy día son el Angoff y sus variaciones, y el Bookmark. Si bien menos reconocidos que los anteriores, otros que tienen desarrollo en la actualidad son el BoW y el de juicio analítico.

Teniendo en cuenta la popularidad y actualidad de los mismos se hará una breve caracterización de cada uno. En un segundo lugar, tomando en consideración los criterios para evaluar los métodos de establecimiento de estándares desarrollados por Berk (1986) se compararán el de Angoff, el BoW, el de juicio analítico y el Bookmark a partir de los aportes de Ricker (2006), Radwan y Rogers (2006), Abbott (2006) y Lin (2006).

### 5.1.1 MÉTODO ANGOFF Y SUS VARIACIONES, EL MÉTODO BOW Y EL MÉTODO DE JUICIO ANALÍTICO

El método Angoff (1971) es uno de los más utilizados y ha recibido varias modificaciones a lo largo del tiempo (Zieky, 2001). En éste se pide a un grupo de panelistas que evalúe cada uno de los ítems de la prueba considerando si un individuo mínimamente competente podría contestarlo de manera correcta o no (Ricker, 2006). El entrenamiento requerido a los jueces para realizar la tarea es mínimo, no es necesario tener a disposición ningún tipo de dato empírico y el procedimiento parece sencillo. En la consigna se les solicita a los panelistas que piensen en un conjunto de 100 estudiantes que presentan un desempeño regular (limítrofe) y que juzguen, para cada ítem de manera independiente, qué porcentaje de esos 100 lo contestaría de manera correcta. En otras palabras, se les pide que establezcan qué probabilidad tiene un estudiante limítrofe de contestar cada ítem de manera correcta[4]. Para cada ítem, se computa entonces un porcentaje o una probabilidad p. Si se suman esos valores p para cada juez se obtiene la cantidad promedio de respuestas correctas

4 "A slight variation of this procedure is to ask each judge to state the probability that the minimally acceptable person would answer each item correctly. In effect, the panelist would think of a number of minimally acceptable persons, instead of only one such person, and would estimate the proportion of minimally acceptable persons who would answer each item correctly" (Angoff 1971, p. 515).

que un estudiante de ese grupo obtendría. Ese será el valor de corte para ese juez. El promedio de los puntajes de corte emitido por el total de los jueces será el puntaje de corte final para la prueba. Un desvío estándar bajo denotará una mayor concordancia entre los jueces.

A pesar de algunos beneficios del método Angoff, los resultados no son siempre satisfactorios (Lin, 2006). Una de sus principales limitaciones es que fue originalmente diseñado para ítems de opción múltiple (selected response items) y no es demasiado permeable a la incorporación de ítems con respuesta abierta (constructed response items).



En segundo lugar, requiere que los jueces evalúen y estimen los valores p para todos los ítems de la prueba, lo cual suele resultar en una tarea sumamente exigente en cuanto al tiempo y los recursos cognitivos que implica (Cizek et al., 2004) y, a su vez, puede resultar en una actividad tediosa para los panelistas (Mitzel, Lewis, Patz & Green, 2001). Algunos estudios han encontrado, incluso, un bajo nivel de precisión en los juicios que realizan los evaluadores para los indicadores de diferente grado de complejidad. En este sentido, Bejar (1983) señala que los panelistas muestran una tendencia a sobreestimar el desempeño de los estudiantes en los ítems difíciles y a subestimar el desempeño en los ítems más sencillos.

Por último, algunos autores aún ponen en duda la capacidad de los jueces para estimar el desempeño de los estudiantes y satisfacer los requerimientos de la metodología Angoff (Lin, 2006). La idea de que alguien pueda tener las capacidades cognitivas para predecir la respuesta de un estudiante limítrofe no resulta, al menos en una primera instancia, algo evidente. Aún más, algunos autores como Shepard, Glaser y Bohrnstedt (1993) lo llegan a considerar impracticable. De acuerdo con Zieky (2001) los jueces podrían a lo sumo llegar a estimar qué tan bien un examinado tendría que responder para obtener un desempeño que sea mínimamente aceptable.

Con el objetivo de generar un mayor grado de acuerdo entre los jueces se han realizado diferentes modificaciones al método Angoff (Ricker, 2006). Algunas de las más comunes implican el uso de algún proceso iterativo que permite a los jueces formular diferentes juicios en cada una de las rondas de testeo y participar de discusiones que den lugar a consensos entre las posturas sostenidas. Otra modificación considera la facilitación de datos normativos y/o estadísticos a los panelistas. Comúnmente, estos datos son brindados antes de la última ronda de discusión. La inclusión de información empírica permite brindar ciertos criterios externos y objetivos que les permita a los panelistas juzgar la relación entre los ítems de distintos grados de dificultad y, además, promover un mayor acuerdo entre los jueces. También algunas modificaciones, como el procedimiento Sí/No (Yes/No method), buscan reducir la complejidad de las tareas solicitadas a los jueces. Tal como se mencionó, originalmente en la consigna del método se les pedía a los panelistas que determinaran la probabilidad de un sujeto limítrofe a contestar de manera correcta un ítem.

Ya a fines de la década del '70 Nassif (1978) y Jaeger (1978) habían sugerido solicitar un juicio por Sí/No en vez de una estimación de probabilidades. De este modo, se simplifica la tarea de los jueces ya que han de evaluar todos los ítems pensando si el estudiante limítrofe contesta de manera correcta o no, evitando la complejidad de probabilidades poco inteligibles.

En el procedimiento Angoff extendido (extended Angoff procedure) sugerido por Hambleton y Plake (1995), se les pide a los jueces que estimen la probabilidad de que un estudiante limítrofe conteste de manera correcta cada ítem, y además, que realicen un ordenamiento

de los ítems según el nivel de dificultad que presentaría para ese mismo "supuesto" estudiante limítrofe. Tal como se puede observar, la tarea es más compleja que en los casos anteriores. No obstante, este procedimiento permite incorporar ítems de respuesta abierta a diferencia de las modificaciones del método Angoff mencionadas.

El método BoW es un método holístico utilizado normalmente con pruebas de respuestas abiertas aunque puede incluir también preguntas cerradas de opción múltiple. Fue descripto inicialmente por Kahl et al. (1994) y ha sido utilizado al evaluar contenidos de los estudiantes como ensayos, textos u otro tipo de producciones similares. Básicamente, el trabajo incluye cinco pasos: (a) establecimiento de los niveles de desempeño; (b) selección de los trabajos de los estudiantes que conformarán una muestra representativa del alumnado a evaluar y categorizar; (c) selección y capacitación de los jueces que participarán del establecimiento de puntos de corte; (d) correspondencia entre el desempeño de los estudiantes y los niveles de desempeño establecidos y; (e) establecimiento de puntos de corte.



En este método, la tarea de los jueces consiste en revisar todas las preguntas y todas las respuestas dadas por los estudiantes que son reunidas en un cuadernillo llamado "cuadernillo de respuestas" y los jueces hacen un análisis global del desempeño de cada estudiante particular indicando en cada cuadernillo el nivel de desempeño alcanzado. Los jueces deben establecer en cada caso cuáles son las habilidades, conocimientos y capacidades observadas en las respuestas justificando de este modo el nivel de desempeño que han indicado.

Usualmente este método incluye tres rondas de juicios. La primera aborda la evaluación de cinco a ocho cuadernillos de respuestas y la revisión de los lineamientos básicos de la metodología. En la segunda, los jueces evalúan una serie de cuadernillos y categorizan cada uno según el nivel de desempeño presentado. Una vez puntuados todos los cuadernillos de respuestas, se eligen los puntajes de corte que reflejen mejor los niveles de desempeño. En la última ronda, se seleccionan otra serie de cuadernillos y se realizan recomendaciones finales acerca de los puntajes de corte obtenidos.

En el método del juicio analítico de Plake y Hambleton (1998, 2001), al igual en el anterior, se les pide a los jueces que clasifiquen la tarea realizada por los estudiantes teniendo en consideración los distintos niveles de desempeño seleccionados. En este caso, la diferencia es que la tarea de los jueces se descompone en partes. En vez de evaluar el material de los estudiantes en su conjunto u holísticamente, los jueces deben evaluar el desempeño de los estudiantes en una tarea por vez. Suponiendo que una evaluación constara de cinco actividades distintas, los jueces deberían realizar una evaluación de la primera actividad en una primera instancia, de la segunda en un segundo momento, y así sucesivamente. Otra diferencia que presenta respecto de los métodos más holísticos es que en este método los jueces no deben poder identificar cuál es el trabajo de cada estudiante en particular. Así como en el método BoW, se utiliza un cuadernillo que reúne todas las respuestas de un mismo individuo, en este caso ante cada evaluación de una actividad debe alterarse el orden de los trabajos de cada estudiante de modo que la evaluación "a ciegas" esté garantizada. El objetivo es que los jueces no puedan reconocer el autor o el estudiante que produce dicha respuesta.

Tal como se puede observar una de las diferencias entre el método de juicio analítico y el método BoW es la cantidad de tiempo que requieren de los jueces. En el primero, al descomponer el trabajo en distintas etapas la tarea suele resultar más fácil y amena, que tener que evaluar todo el conjunto de trabajo en una única instancia como sucede en el método BoW.

## 5.1.2 COMPARACIÓN DEL MÉTODO ANGOFF, EL MÉTODO BOW, EL MÉTODO DE JUICIO ANALÍTICO Y EL MÉTODO BOOKMARK SIGUIENDO LOS CRITERIOS ENUNCIADOS POR BERK (1986)

En el siguiente capítulo se desarrollará en extenso las características de la metodología Bookmark. No obstante, a los fines de comparar la metodología Angoff, el método BoW, el método de juicio analítico y el método Bookmark, hace falta mencionar –al menos sucintamente– algunos aspectos relevantes de este último.

El método Bookmark, a diferencia de los otros procedimientos mencionados supone algunas ventajas ya que permite: (a) el uso de ítems de respuestas abiertas y respuestas cerradas; (b) simplifica las tareas cognitivas requeridas a los jueces que deben establecer los puntajes de corte y; (c) integrar los contenidos de la pruebas a determinados descriptores de niveles de logros (Mitzel et al., 2001).

En los artículos de Ricker (2006), Radwan y Rogers (2006), Abbott (2006) y Lin (2006) las metodologías Angoff, BoW, de juicio analítico y Bookmark son evaluadas utilizando los criterios propuestos por Berk (1986) para los métodos de establecimiento de estándares. De ese modo, se analizan de manera crítica las debilidades y fortalezas de cada uno de los procedimientos. En los cuatro casos, de manera independiente, se utiliza una escala de tipo Likert de 3 puntos (1 = no se ajusta; 2 = se ajusta parcialmente; 3 = se ajusta complemente) con el objetivo de evaluar qué tan bien se ajusta cada método a los términos enunciados. En la tabla 1 se pueden observar las puntuaciones brindadas por los autores en función de una exhaustiva evaluación de cada uno de los métodos. En resumen, el método Bookmark parece ofrecer mayores beneficios que la metodología Angoff y el resto de las metodologías revisadas.

| METODOLOGÍA | | | | | |
|---|---|---|---|---|---|
| **Criterios de Berk** (1986) | | **Angoff** (Ricker, 2006) | **Bookmark** (Lin, 2006) | **BoW**(Radwan & Rogers, 2006) | **Juicio analítico** (Abbott, 2006) |
| De adecuación técnica | Clasificación apropiada de los datos | 2 | 3 | 3 | 3 |
| | Sensible al desempeño de los examinados | 2 | 3 | 3 | 2 |
| | Consideración de aspectos relacionados con la instrucción y entrenamiento de los jueces | 1 | 3 | 3 | 3 |
| | Aporte de datos y sustento estadístico para la toma de decisiones | 3 | 3 | 2 | 3 |
| | Identificación del "verdadero" estándar | 2 | 3 | 3 | 1 |
| | Aporte de evidencias de validez para las decisiones | 2 | 2 | 1 | 1 |
| Factibilidad | Fácil de implementar | 2 | 3 | 2 | 2 |
| | Fácil de computar | 3 | 3 | 3 | 3 |
| | Fácil de interpretar | 3 | 3 | 2 | 3 |
| | Confiable para los no especialistas en el área | 3 | 3 | 2 | 2 |

**Tabla 1:** Comparación entre la metodología Angoff, BoW, juicio analítico y Bookmark siguiendo criterios de Berk (1986)

Algunos autores han sintetizado las ventajas y beneficios del método Bookmark. Entre las fortalezas se destacan: (a) poder trabajar con ítems de respuesta abierta y respuesta cerrada; (b) reducir el esfuerzo cognitivo requerido a los panelistas; (c) incluir tareas sencillas para los jueces; (d) promover un mejor entendimiento de lo que se espera como "buen desempeño académico"; (e) eficiencia en el establecimiento de los puntos de corte; (f) capacidad para poder integrar datos estadísticos a la evaluación y juicio de los panelistas; (g) uso eficiente del tiempo y bajo nivel de errores en el establecimiento de los puntajes de corte (Herrera Ortiz et al., 2009; Lin, 2006).



## 5.2 EL SUSTENTANTE LIMÍTROFE

Todo proceso de establecimiento de estándares contiene elementos ligados a la subjetividad y la arbitrariedad –particularmente aquellos que dependen de la formulación del juicio de experto–. En cualquiera de los métodos para el establecimiento de los puntos de corte, lo que se busca en última instancia es minimizar la arbitrariedad y la subjetividad de los resultados obtenidos. Uno de los elementos clave para ello es una adecuada definición del sustentate limítrofe, también llamado: sustentante de competencia mínima, sustentante límite o sustentante borderline.

El concepto de sustentante limítrofe ha de ser claro y común para todos los panelistas que participen (Zieky, Perie & Livingstone, 1986). En caso contrario, se corre el riesgo de que las concepciones subjetivas de los panelistas influyan sobre los puntajes de corte establecidos. A este respecto, investigaciones previas han evidenciado que los puntajes de corte suelen presentar un rango de variación más amplio entre los panelistas cuando el sustentante limítrofe no se encuentra claramente definido de antemano (Impara et al., 2000).

Poder distinguir cuando un estudiante se encuentra en el límite de cualquiera de las categorías de desempeño establecidas (e.g. bajo, medio y alto) forma parte de cualquier procedimiento. En consecuencia, diferentes autores recomiendan que los jueces puedan ponerse de acuerdo acerca de cuáles serán las habilidades y conocimientos mínimos requeridos para cada categoría o para cada nivel antes de proceder a establecer los puntos de corte (Nassif, 1978). Livingston y Zieky (1982) recomiendan que en el proceso de establecimiento de puntos de corte haya un espacio donde los participantes describan con sus propias palabras cuáles serían las características de un estudiante limítrofe. En particular, qué conocimientos y habilidades debería y no debería presentar.

Teniendo en cuenta que los descriptores de cada nivel de logro pueden ser redactados antes o después del establecimiento de puntos de corte, el momento en que se disponga para esta actividad también influirá sobre la profundidad con que se debatan las capacidades y conocimientos mínimos que se supone han de presentar los individuos límites entre niveles. En el caso en que los descriptores sean redactados de manera previa, los participantes deberán hacerse una idea precisa y clara de las características que ha de reunir el sustentante límite antes de realizar las tareas relativas al establecimiento de puntajes de corte. En el caso de que los descriptores sean redactados posteriormente, existe una inherente permeabilidad en lo que refiere a la conceptualización del sustentante límite, dado que en este último caso los descriptores y, por tanto las capacidades, habilidades y contenidos que deben presentarse en cada categoría de desempeño, es delimitada al final de todo el proceso de establecimiento de estándares.

# 6. MÉTODO BOOKMARK



Como se ha mencionado anteriormente, el establecimiento de estándares comprende cuatro tareas básicas: (1) selección de la cantidad de niveles de desempeño que se desean establecer; (2) elección de los nombres o etiquetas de cada nivel; (3) redacción de los descriptores de cada nivel y; (4) establecimiento de los puntajes de corte correspondientes.

Habiendo revisado la bibliografía publicada en relación a los diferentes métodos disponibles y reunido evidencia suficiente que acredita la idoneidad del método Bookmark, se seleccionó este método para la determinación de los puntos de corte de los distintas disciplinas y años evaluados en las escuelas primarias y secundarias de la República Argentina que participaron del dispositivo Aprender durante octubre de 2016.

## 6.1 DESCRIPCIÓN DEL MÉTODO

El método Bookmark fue introducido por Lewis et al. en Estados Unidos en 1996 con el objetivo de evitar las limitaciones asociadas a los métodos de determinación de estándares previos (Karantonis & Sireci, 2006). Específicamente, implica una serie de procedimientos y actividades diseñadas para el estableciendo de los puntos de corte –que son las operacionalizaciones numéricas de las descripciones cualitativas de los niveles de desempeño– (Cizek & Bunch, 2007; Herrera Ortiz et al., 2009). Como todo método que se basa en el juicio de experto, este método se asienta en la revisión cuidadosa que realiza un grupo de especialistas. En el método Bookmark esta revisión es efectuada sobre un conjunto limitado de los ítems que compone la prueba, que son presentados en un cuadernillo siguiendo un orden de dificultad creciente desde lo más sencillo a lo más complejo.

Una de las principales características de este método es que ha sido diseñado utilizando la Teoría de Respuesta al Item (TRI). En este sentido, las pruebas basadas en la TRI otorgan algunos beneficios a sus usuarios. En primer lugar, gracias a la propiedad de invarianza de los parámetros de los procedimientos de la TRI, una vez establecidos los puntos de corte adecuados, éstos no variarán aunque se modifique o añadan nuevos reactivos al banco de datos –siempre que se utilicen los diseños de equiparación métrica necesarios– (García, Abad, Olea & Aguado, 2013). Además, la TRI permite mapear los ítems según su nivel de dificultad. Es decir, permite calibrar los ítems para presentarlos ordenados en dificultad a los jueces que participarán en el establecimiento de puntos de corte. Además, este ordenamiento puede ser realizado utilizando diferentes indicadores –de respuestas abiertas o de construcción de respuesta y de respuesta cerrada o de opción múltiple–, logrando simplificar en gran medida las tareas requeridas a los jueces (Lewis, Green, Mitzel, Baum & Patz, 1998). De ahí que otra característica distintiva del método es la utilización de material externo que facilita el trabajo a realizar y que constituye una herramienta básica para que los jueces puedan tomar sus decisiones (Herrera Ortiz et al., 2009). Los ítems se presentan en un Cuadernillo de Items Ordenados (Ordered Item Booklet – OIB, por sus siglas en inglés) y la dificultad de cada indicador es determinada empíricamente.

El cuadernillo permite a los panelistas tener una visión global del grado de complejidad de la prueba y ayuda a que focalicen su atención en aquellos ítems particularmente sensibles para los estudiantes limítrofes. En el OIB se presenta un ítem por página. A los jueces se les solicita que fijen una marca para cada punto de corte. Es decir, para tres niveles de desempeño se colocarán dos puntos de corte o dos marcas. Cada marcador es colocado entre dos indicadores límites de manera que separe aquellos estudiantes que se espera tengan un mínimo dominio de los indicadores previos, pero no han logrado aún dominar los elementos posteriores. Específicamente, los jueces deben "determinar cuáles son los reactivos que un sustentante límite podría contestar correctamente con una probabilidad mayor de .67" (Herrera Ortiz et al., 2009, p. 45). Si el juez considera que para un ítem la probabilidad de que el individuo límite conteste de manera correcta es menor, deberá colocar el marcador en ese mismo punto señalando el corte entre las categorías o niveles de desempeño previamente establecidos. En la figura 1 se puede observar una gráfica del procedimiento enunciado.

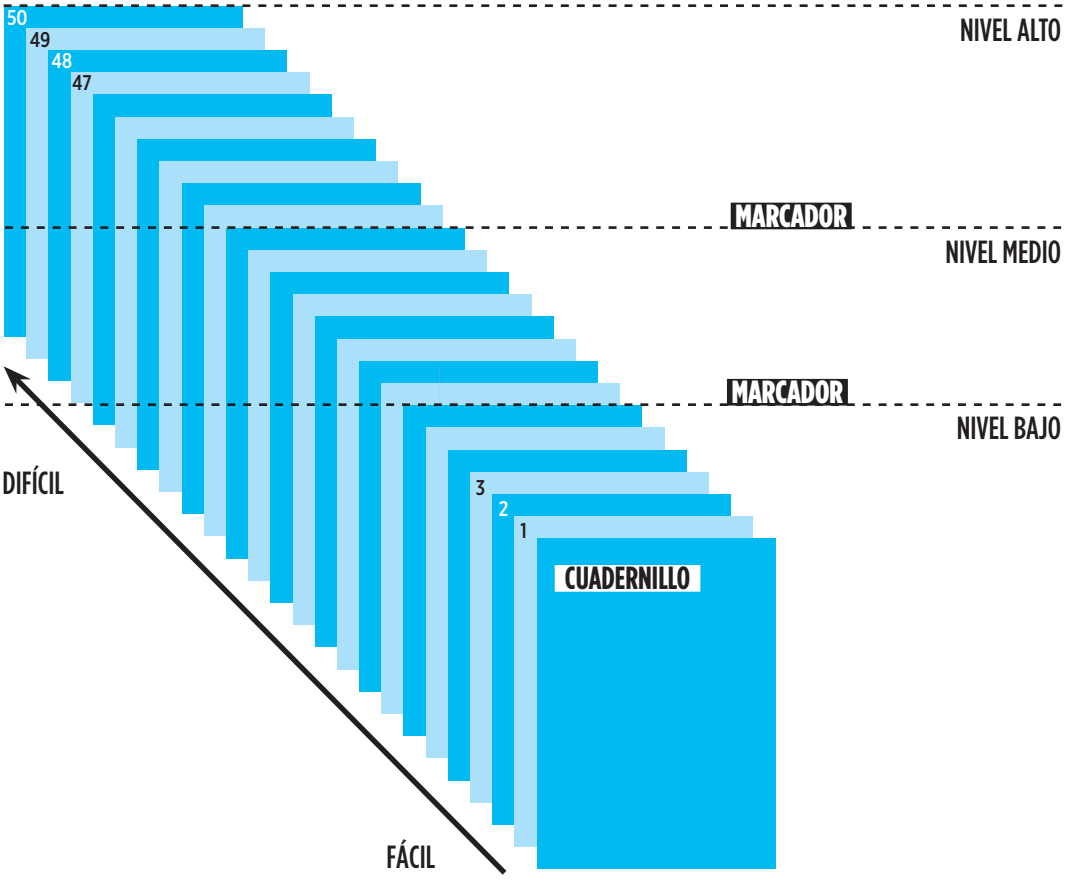


**Figura 1.** Cuadernillo de Ítems Ordenados para el método Bookmark.
**Nota.** Adaptado de Mitzel et al. (2001), p. 253.

Durante la evaluación de los ítems para la colación del marcador, puede suceder que algunos participantes objeten la ubicación de algún ítem o algunos ítems en particular (Skaggs & Tessema, 2001). Pueden hacer alguna observación y disentir sobre el ordenamiento del cuadernillo o el nivel de dificultad de ese ítem. De acuerdo con Lewis y Green (1997) existen dos explicaciones para estas divergencias de criterio: (1) las diferencias que pueden surgir en la implementación de los currículos entre jueces que pertenecen a distintos distritos o localizaciones; (2) la dificultad de los jueces para estimar de manera correcta la dificultad del ítem.



Skaggs y Tessema (2001) señalan otros factores que pueden influir en las diferencias entre la percepción de los jueces y el ordenamiento basado en la TRI. En primer lugar, un error en la estimación de los parámetros de los ítems podría alterar su ordenamiento. En segundo lugar, los autores señalan que los jueces podrían estar ignorando algunas características de los ítems que afectan su dificultad. Por ejemplo, en los ítems de respuesta cerrada de opción múltiple la dificultad del ítem no sólo está dada por su contenido o por cómo ha sido redactado sino también por las alternativas de respuestas que lo acompañan. En este sentido, la facilidad para adivinar una respuesta o la cantidad de distractores presentes en las alternativas de respuestas también son factores que añaden más o menos complejidad al reactivo. Una última razón que señalan Skaggs y Tessema, en particular referidas a los exámenes de lengua inglesa, es que algunos ítems combinan el nivel de dificultad de los textos de lectura evaluados y las habilidades necesarias para comprender esos pasajes.

Muchos de los panelistas muestran desacuerdo en el ordenamiento de ítems que corresponden a textos más difíciles y se presentan al inicio del cuadernillo, mientras que ítems que corresponden a textos más fáciles se ubican posteriormente. La dificultad del texto influye sobre la percepción que tienen los docentes de la dificultad del ítem. De ahí, la disonancia.

Por su parte, Jornet Meliá y Bakhoff (2006) observan que este tipo de discrepancias también puede darse por: (a) un contenido teóricamente simple evaluado por un ítem mal diseñado de manera que eso hace que aparente ser más difícil; (b) el caso contrario, un contenido teóricamente complejo evaluado por un ítem mal diseñado, de manera que en la redacción del ítem se incluyen pistas que orientan hacia la identificación de la respuesta correcta y; (c) un contenido que no se enseñe habitualmente en el aula aunque esté presente en el currículo (por ejemplo, los autores señalan los contenidos de Estadística en cursos de Matemática).

Como una solución frente a este problema, Lewis y Green (1997) sugieren dar lugar a una discusión detallada de qué mide el ítem en cuestión y por qué muestra una mayor complejidad que el ítem anterior. De acuerdo con los autores, esta estrategia sería útil para resolver estos obstáculos y avanzar con el trabajo.

## 6.2 ANTECEDENTES DE SUS APLICACIONES

En la actualidad el método Bookmark es el más utilizado en la mayoría de los estados norteamericanos (Nellhaus, 2000). Diferentes países de Centro América y Latinoamérica también han reportado resultados satisfactorios en su implementación. El método ha sido utilizado en Perú, Chile, Guatemala, Brasil y México en diferentes disciplinas y grados académicos. Además, la cantidad de puntos de corte seleccionados, así como los niveles de logro y sus respectivas etiquetas también varían en cada país según los objetivos de la evaluación.

Existen diferentes informes técnicos acerca de las características de la implementación del método Bookmark en varios estados de EE.UU. Es necesario observar que, a diferencia de los países mencionados y de la Argentina, en EE.UU. cada estado aplica su propia metodología para el establecimiento de estándares. Algunos de los estados que han aplicado la metodología Bookmark son Wisconsin (Department of Public Instruction & CTB/McGraw-Hill, 2003), Kansas (Perie & Smith, 2015) e Indiana (Egan, Barton & Roeber, 2015).

En Wisconsin se establecieron los puntajes de corte de Lectura, Artes, Matemáticas, Ciencia y Ciencias Sociales de seis grados (3°, 4°, 7°, 8°, 9°y 10°). Se determinaron cuatro niveles de desempeño cuyas etiquetas fueron: Mínimo, Básico, Competente y Avanzado; y tres puntos de corte. Participaron un total de 240 panelistas de 84 escuelas. Los concurrentes fueron nominados por organizaciones de profesionales o por otras fuentes de referencias o, incluso, autonominados. Se le solicitó a cada nominado que aportara sus calificaciones y experiencias. Un comité eligió los mejor cualificados para realizar el trabajo.

En tanto, en Kansas se evaluaron las disciplinas de Lengua Inglesa y Artes, y Matemáticas en siete grados escolares (Perie & Smith, 2015). Participaron un total de 160 maestros y la duración del trabajo fue de cuatro días. Se establecieron cuatro niveles de desempeño y tres puntos de corte para cada caso.

En Indiana, también se evaluaron las disciplinas de Lengua Inglesa y Artes, y Matemáticas en seis grados (Egan et al., 2015). Los participantes fueron seleccionados por el Departamento de Educación con el propósito de ser representativos de las distintas regiones geográficas, localización de las escuelas (urbana, suburbana, rural) y nivel socioeconómico. En relación a la zona geográfica 38% pertenecía al norte, 42% al centro y 20% al sur. En relación a la localización de la escuela 31% se ubicaba en el sector urbano, 29% en el sector suburbano y 40% en el rural. Por último, 77% era de un nivel socioeconómico medio y 23% de nivel de socioeconómico bajo. También se seleccionaron los participantes en función de sus años de experticia. El 12,5% tenía entre 1 y 5 años de experiencia, el 14,3% tenía entre 6 y 10, el 17% tenía entre 11 y 15, el 19,6% tenía entre 16 y 20 y el 35,7% tenía más de 20 años de experiencia en su trabajo. Un total de 110 panelistas participaron en el procedimiento que duró cuatro días. De los 110 panelistas se seleccionaron 4/5 líderes por grupo (es decir entre 24 y 30 líderes en total) para que coordinaran los debates en cada una de las rondas de trabajo.

Dentro de los trabajos realizados en nuestra región, se encuentra el caso de Chile donde se aplicó la metodología Bookmark para determinar los puntajes de corte de 4° grado básico en las áreas de Lectura, Matemática, compresión del Medio Natural y Comprensión del Medio Social y Cultural (Comisión para el Desarrollo y Uso del Sistema de Medición de la Calidad de la Educación [SIMCE], 2006). Se establecieron dos niveles de logros (Nivel de logro Intermedio y Nivel de logro Avanzado) y un puntaje de corte. Participaron en el estudio 123 panelistas (Matemática = 28; Lectura = 27; Comprensión del Medio Natural = 34; Comprensión del Medio Social y Cultural = 34) entre quienes se encontraban profesores de aula de primer ciclo básico, académicos de universidades, especialistas de centros de investigación, especialistas de congregaciones religiosas, especialistas del Ministerio de Educación de distintas regiones (región metropolitana, de la zona sur y del norte del país), docentes de establecimientos municipalizados, particulares subvencionados y particulares rentados. Una vez obtenidos los puntajes de corte para cada una de las disciplinas, se construyó un intervalo de confianza que fue presentado a un comité técnico con la finalidad de definir un puntaje de corte ubicado dentro del rango recomendado por los especialistas.

Por otra parte, el trabajo realizado en Perú constó de una evaluación previa en la que se comparó los resultados de la metodología Bookmark y la del método canasta para poder analizar cuál de estos procedimientos arribaba a mejores resultados (Cruz Ampuero, Espinoza Pezzia, Montané Lores, &, Rodríguez Cuellar, 2001). Una vez contrastados ambos métodos se optó por la aplicación del método Bookmark en la evaluación de dos áreas disciplinares: Lógico Matemática y Comunicación Integral. Se establecieron dos puntos de corte y tres niveles de logro (Básico, Suficiente y Avanzado), participando de este proceso un grupo de 8 jueces para el área de Lógico Matemática y 12 para el área de Comunicación Integral. Los panelistas eran docentes de cada área y grado evaluado y fueron cuidadosamente seleccionados teniendo en cuenta el dominio de su área, los años de experiencia como docentes, su nivel educativo, entre otros aspectos. Además se buscó que los grupos fueran representativos de diferentes tipos de gestión (estatal, privada, parroquial), niveles socioeconómicos y culturas socio-ambientales (rural y urbana).



En el caso de Guatemala, el método Bookmark fue utilizado para poder establecer los puntajes de corte para el área de Lectura en español y Matemática (Ical Choc et al., 2009). Se seleccionaron tres puntos de corte y cuatro niveles de logro (Insatisfactorio, Debe mejorar, Satisfactorio, Excelente). La muestra estuvo compuesta por 45 panelistas entre quienes se incluían docentes, maestras especiales, docentes bilingües, administrativos, expertos en currículo, miembros de consejos escolares, maestras de otras áreas curriculares y padres. Los jueces fueron convocados con anticipación por correo postal.

## 6.3 MATERIALES Y PROCEDIMIENTOS DEL MÉTODO BOOKMARK

### 6.3.1 CUADERNILLO BOOKMARK Y CUADERNILLO COMPLEMENTARIO

El cuadernillo Bookmark muestra los reactivos ordenados por nivel de dificultad creciente, de lo más simple a lo más complejo. La utilización de este cuadernillo busca facilitar el trabajo de los evaluadores promoviendo que puedan tener una idea global e integrada de lo que mide la prueba, así como también servir como instrumento para la determinación adecuada de los puntos de corte (Lewis et al., 1998). Estos autores sugieren incluir un formulario de mapeo de los ítems junto con los cuadernillos, como una guía para trabajar con el cuadernillo. Este formulario enumera todos los ítems tal y como aparecen en el cuadernillo y aporta información relativa a cada ítem: ubicación en la escala, ubicación en la prueba operativa (operational test), el estándar u objetivo al que apunta el ítem, y un espacio para que los panelistas hagan sus comentarios acerca de qué mide cada ítem y por qué le parece más difícil que el ítem precedente, así como también los puntajes de corte colocados para cada una de las rondas de evaluación (Lewis et al., 1998).

El cuadernillo puede incluir un mayor número de ítems que los considerados en la prueba administrada a los estudiantes. En este sentido, pueden tomarse reactivos que no fueron evaluados en la versión del examen pero sí se encuentran en el banco de ítems (Herrera Ortiz et al., 2009; Mitzel et al., 2001). Idealmente, el cuadernillo debe contener una cantidad de reactivos suficientes para representar el continuo de dificultad. Se desaconseja que existan franjas amplias de dificultad no representadas por ningún reactivo. Por ejemplo, si en un examen ya administrado se tienen dos reactivos con un índice de dificultad (en cierta escala) de 1.05 y 1.25 y entre estos niveles de dificultad no existe ningún reactivo, para el cuadernillo se podrían incluir reactivos del banco de ítems con índices de dificultad ubicados en esa franja. Por ejemplo de 1.10, 1.15 y 1.20. De esta manera el continuo de

dificultad es mejor cubierto y la determinación del punto de corte puede realizarse de una manera más precisa.

Siguiendo las recomendaciones de Lewis, Mitzel, Mercado y Schulz (2012), se sugiere además que el cuadernillo reúna al menos entre 40 y 50 reactivos. De acuerdo con los autores, tener suficientes ítems en el cuadernillo es importante por dos motivos. Por un lado, garantiza que los participantes dispongan de un amplio rango de referencia para poder realizar sus juicios. Cuanto mayor es la evidencia de la que disponen los jueces acerca de los contenidos y habilidades representados por los ítems, mejor. Por ello, a mayor cantidad de ítems se supone que hay una mejor cobertura y puede lograrse un conjunto representativo de todas las capacidades y contenidos evaluados. Por otro lado, un número menor de ítems puede dar lugar a un mayor intervalo en los niveles de dificultad de un ítem a otro y esto afectaría la precisión de los puntos de corte. No hay investigaciones previas que estudien la mínima cantidad de ítems necesarios para llevar a cabo el procedimiento, pero sí es de común acuerdo que a mayor cantidad de puntos de corte, mayor cantidad de ítems necesarios, dado que deben ser representados más contenidos y habilidades. Sintetizando, no se recomiendan menos de 40 a 50 ítems por cuadernillo para trabajos donde se deban estimar dos puntos de corte. Sin embargo, también se ha de tener en cuenta la cantidad de niveles de desempeño. En este sentido, a mayor cantidad de puntos de corte, mayor cantidad de ítems que deberá reunir el OIB (Karantonis & Sireci, 2006).

Tal como se mencionó anteriormente, el cuadernillo presenta un ítem por hoja. En la esquina derecha de la hoja se muestra la ubicación del ítem en el OIB: 1 para el más simple, 2 para el siguiente y así consecutivamente. Es importante que este número sea claramente identificable del resto de los números ubicados en la hoja, ya que será el número que los jueces tengan que informar cuando establezcan el punto de corte. Puede presentarse en un tamaño de fuente mayor y en negrita de modo que sobresalga. En la esquina izquierda de la hoja se muestra la ubicación del ítem en el examen administrado. El ítem debe aparecer tal y como fue administrado a los estudiantes, incluyéndose las alternativas de respuesta y señalándose con un asterisco cuál es la respuesta correcta. También se puede indicar el nivel de habilidad necesario para responderlo de manera correcta con una probabilidad de .67 (Mitzel et al., 2001).

En el caso de algunas disciplinas como Lengua o Geografía muchas veces los ítems se encuentran acompañados con pasajes de texto, figuras o mapas. Los textos o figuras que acompañen los ítems deben ser mostrados en un cuadernillo aparte. Este cuadernillo recibe el nombre de cuadernillo complementario (companion booklet). Es recomendable que el mismo sea testeado previamente para comprobar que no presente errores y se encuentre bien impreso. La idea de incluir esta información en un cuadernillo complementario sirve a los fines de facilitar la manipulación de los materiales a los jueces. Cada ítem del OIB muestra en un recuadro aparte en qué página del cuadernillo complementario puede hallar el panelista, la información o estímulo correspondiente a ese ítem.

### 6.3.2 PROBABILIDAD DE RESPUESTA

La tarea fundamental de los jueces en el método Bookmark es determinar cuáles son aquellos reactivos que un sustentante límite podría contestar correctamente con una probabilidad mayor de .67 (el valor 0.67 está relacionado con que dos de cada tres estudiantes puedan resolver el ítem, $2/3 \approx 0.67$). Esta tarea es comúnmente formulada mediante la

siguiente pregunta: "¿Qué tan probable es que un estudiante limítrofe conteste este ítem de manera correcta?". Sin embargo, a esta pregunta hace falta agregarle un elemento más: el nivel de probabilidad.

El nivel de probabilidad es uno de los pilares del método ya que su estimación se relaciona con la cantidad de parámetros utilizados por cada modelo logístico. El nivel de probabilidad de .67 es el nivel sugerido por los especialistas en el área, debido a que con él se maximiza la función de información de una prueba al utilizar un modelo de dos parámetros (Huynh, 2000). El objetivo de los jueces es entonces revisar el cuadernillo y, pensando en el individuo limítrofe, indicar si la probabilidad de contestar correctamente cada ítem es mayor a 0,67. A veces, este nivel puede reemplazarse por la fracción 2/3. Matemáticamente son similares y algunos estudios han evidenciado que a las personas les resulta más fácil pensar en términos de frecuencias que en términos de probabilidades (Gigerenzer & Hoffrage, 1995).



En las actividades de capacitación y entrenamiento del método dirigidas a los panelistas se suele sugerir: "Piense en un grupo representativo de estudiantes que se encuentre en el límite superior del nivel bajo. Dos de cada tres de estos estudiantes, ¿contestaría este ítem de manera correcta?". La idea es ir presentando la información dosificada. Una alternativa sería pedirle a los jueces que hagan una lectura de los ítems del más fácil al más difícil: "Por favor, vaya recorriendo los ítems del más fácil al más difícil. Al hacer esta lectura: ¿cuál es el primer ítem que un estudiante de nivel bajo probablemente no respondería bien? Piense en un estudiante que se encuentre en el borde del nivel bajo".

La ubicación del marcador determina que a partir de ese ítem, el individuo limítrofe tiene una probabilidad menor a 2/3 de contestar correctamente los reactivos. O a la inversa, implica que el sustentante límite tiene una probabilidad mayor de 0,33 de errar en su respuesta. En otras palabras, esto significa que el estudiante limítrofe todavía tiene alguna posibilidad de contestar correctamente un ítem posterior, pero que esa probabilidad es muy baja (Jornet Meliá & Backhoff, 2008). El marcador se colocará entonces en aquel indicador que dos de cada tres sustentantes limítrofes no puedan contestar de manera correcta (Cizek et al., 2004; Cizek & Bunch, 2007).

Cada juez realiza su tarea de manera individual. Una vez que establece su marcador se discute qué es lo que hace que un determinado reactivo sea más difícil que aquellos que lo preceden (Herrera Ortiz et al., 2009). Luego de haber establecido el primer marcador, los panelistas deberán seguir analizando el resto de los ítems correspondientes al siguiente nivel y punto de corte. En esta segunda actividad, no será necesario revisar los ítems precedentes al primer marcador. En el primer caso, los jueces deberán focalizar su atención en el individuo limítrofe entre el nivel de desempeño Bajo- Medio, y en el segundo caso, entre el nivel de desempeño Medio-Alto.

### 6.3.3 ESTIMACIÓN DE LOS PUNTOS DE CORTE

El paso siguiente en el procedimiento implicará la obtención de los puntos de corte. En este sentido es preciso aclarar que los marcadores no se corresponden en forma directa con los puntos de corte del examen. Por ejemplo, si un juez tuviera un cuadernillo de 60 ítems y colocara el marcador en el ítem 25, el ítem 25 no equivaldría al punto de corte. Para determinar los puntos de corte se requiere poder calcular la habilidad (puntaje TRI) necesaria para acertar ese ítem marcador con una probabilidad de 0.67.

### 6.3.4 CAPACITACIÓN DE LOS PANELISTAS

La capacitación de los panelistas es uno de los pasos fundamentales del método dado que el resultado dependerá del entendimiento que los jueces tengan de su tarea. Es importante que los participantes comprendan el trabajo a realizar y su papel dentro de las actividades planteadas. El entrenamiento también debe orientarse a involucrar a los jueces en las tareas y a motivarlos de modo que se genere un clima facilitador. Para ello también resulta beneficioso informar y dejar en claro cuál es el fin último del procedimiento de establecimiento de estándares ya que de esta manera se provee un marco de referencia al trabajo (Cizek et al., 2004).

Básicamente, el entrenamiento estará orientado a: (a) explicar los pasos a seguir para el establecimiento de estándares; (b) presentar y familiarizar a los panelistas con los documentos en donde deberán establecer sus puntajes de corte; (c) completar una versión "de prueba" o "piloto" de los formularios de evaluación; (d) practicar la determinación de un punto de corte; (e) explicar cualquier tipo de dato estadístico que será utilizado en el proceso; (f) familiarizar a los panelistas con el contenido de las pruebas; (g) repasar el conjunto de ítems sobre el cual se aplicará el procedimiento y/o en algunos casos, completar un cuadernillo de prueba (Hambleton, 2001).

Durante el entrenamiento también se deberá proveer información acerca de las etiquetas y niveles de desempeño que se desean establecer (Lewis et al., 1999). De acuerdo al plan de cada institución u organización, los descriptores podrán ser discutidos y redactados por los mismos panelistas antes del establecimiento de los puntos de corte, o posteriormente. Incluso, se puede brindar a los panelistas algunos descriptores guía que luego serán ajustados al finalizar el trabajo (una vez determinados los puntos de corte). Las descripciones guía muestran a los jueces de una manera global "qué tan rigurosos serán los estándares para describir los niveles de logro y para ello se recomiendan palabras o conceptos que marquen una clara diferencia entre las categorías" (Herrera Ortiz et al., 2009). En el método Bookmark es común que los descriptores no se encuentren totalmente definidos hasta que los puntos de corte son establecidos (Mitzel et al., 2001). Una vez definidos los puntos de corte para cada categoría, entonces sí, suele describirse en detalle cuáles son los conocimientos, habilidades y capacidades esperados para cada categoría.

Además, en la capacitación se sugiere explicitar sobre qué grupo se aplicarán los estándares y cuáles son las características de ese grupo. En relación a ello, también se recomienda generar un espacio de debate acerca de las características del sustentante límite. Esto permite que haya un espacio previo de reflexión sobre este aspecto y que los jueces reconozcan y acuerden las características del individuo limítrofe (Cizek et al., 2004). En el método Bookmark será necesario que los panelistas puedan formarse una idea respecto de en qué punto los estudiantes de una determinada categoría (e.g. Bajo) tendrán una probabilidad especifica (e.g. 0.67) de responder correctamente.

Por otra parte, en algunos casos se suele prevenir informar a los jueces acerca de factores que pueden incidir en el ordenamiento de los ítems en el cuadernillo y sobre otros externos que pueden impactar en los resultados de las pruebas. Es sabido que en las evaluaciones intervienen circunstancias externas por lo que es conveniente hacer una puesta en común con los panelistas antes de dar comienzo a su trabajo. En este sentido, se mencionan como factores influyentes: (a) el tiempo limitado para realizar la prueba; (b) la dificultad que presentan los ítems de opción múltiple (a veces algunas respuestas incorrectas –distracto-

res– pueden ser muy similares a las correctas); (c) la artificialidad inherente a los sistemas de evaluación; (d) el papel que juega la tendencia de los estudiantes a "adivinar" la respuesta o marcar una cualquiera, cuando se trata de un ítem de opción múltiple, entre otros.

En esta etapa también se aconseja notificar a los participantes acerca de las modificaciones que podrán ser realizadas a posteriori por el equipo técnico y/o asesores políticos sobre las puntuaciones corte. Al respecto, se ha de precisar que en algunas ocasiones los puntos de corte finales no se mantienen idénticos a los obtenidos en las sesiones de trabajo. Esto se debe a que cada grupo de panelistas trabaja sobre un área de contenido específico (e.g. Lengua, Matemática, Ciencias Sociales) o sobre un grado específico (e.g. tercer grado, cuarto grado). A veces, autoridades de determinadas agencias, secretarias o ministerios deben realizar cambios en los puntajes de corte establecidos para poder articular la información brindada por los distintos grupos de panelistas. Entonces, resulta recomendable explicar estas cuestiones en la capacitación, de modo que el proceso sea transparente y ninguno de los participantes vea su trabajo minusvalorado. La capacitación también debe prever la selección y contacto de quienes se desempeñarán como líderes en los paneles de evaluación. En el siguiente apartado se hará mención a este aspecto.

Por último, una vez finalizada la capacitación se recomienda evaluar la comprensión de las tareas a través de una encuesta (Herrera Ortiz et al., 2009). En la figura 2 se muestra un ejemplo de cuestionario autoadministrable.

| | |
|---|---|
| | **EXAMEN DEL ÁREA:** |
| | **GRADO EVALUADO:** |
| | **NOMBRE DEL PANELISTA:** |
| ○ | **INSTRUCCIÓN:** A continuación se presentan algunas preguntas que hacen referencia a las actividades realizadas durante la capacitación en el método Bookmark. Por favor, lee cada una con atención. No hay respuestas correctas o incorrectas. Marca la opción que represente mejor tu modo de pensar.<br><br>**A.** He comprendido las tareas que se deberé realizar en los siguientes días<br>☐ Sí ☐ No |
| ○ | **B.** He comprendido cuál es la pregunta que debo responder en cada una de las rondas de evaluaciones<br>☐ Sí ☐ No |
| | **C.** He comprendido la repercusión que tendrán los puntos de corte establecidos en este trabajo<br>☐ Sí ☐ No |

**Figura 2.** Ejemplo de formulario para evaluar las actividades de capacitación. Nota. Adaptado de Herrera Ortiz et al. (2009), p. 55.

### 6.3.5 COMPOSICIÓN DEL PANEL DE PARTICIPANTES

En el apartado Validez y confiabilidad de los métodos de establecimiento de estándares (ver p. 21) se hizo referencia a la importancia de la composición de la muestra para la validez y confiabilidad de los resultados obtenidos. Cuanto más representativa es la muestra de todos los actores sociales que participan en el ámbito educativo y cuánto más representativa de las diferentes escuelas, mayor será la confiabilidad de los resultados.

De acuerdo con Hambleton (2001) la selección de la muestra deberá ser intencional siguiendo un criterio racional. El número de participantes que componga la muestra debe ser considerado con cuidado, teniendo en cuenta que debe lograrse una adecuada representatividad en cuanto a la ubicación geográfica, nivel socioeconómico, género, edad, tipo de escuela a la que pertenecen (urbanas, suburbanas, rurales) y la experiencia laboral y/o académica (Lewis et al., 1999). Es recomendable que antes del proceso de establecimiento de estándares se les pida a los participantes que completen un cuestionario de datos sociodemográficos para poder contar con estos datos.

Se sugiere que cada panel se encuentre integrado por al menos 18 participantes (para cada grado o disciplina evaluada). Idealmente, se recomienda un mínimo de 24 (Lewis et al., 1999). Además, los panelistas que participen de la determinación de los puntos de corte deberán ser expertos en su área, mostrar un manejo exhaustivo de los contenidos y habilidades examinados, y conocer el grupo que está siendo evaluado (Impara & Plake, 1997). También se recomienda que en la capacitación los jueces tomen contacto con el examen administrado, los temas y conocimientos incluidos, el objetivo del examen y cómo se darán a conocer los resultados (Cizek & Bunch, 2007).

Usualmente, el plantel completo es dividido en tres o cuatro grupos pequeños de manera de permitir una mayor discusión entre los panelistas. Cada grupo contiene entre 5 a 7 miembros.

Líderes de los pequeños grupos. El entrenamiento de los líderes de cada grupo pequeño involucra una revisión del cronograma para la determinación de los estándares y un asesoramiento acerca de las tareas que le son encomendadas tales como facilitar la discusión grupal, mantener la atención del grupo en el tema a tratar y supervisar el tiempo del que disponen para cada una de las discusiones (Lewis et al., 1999).

### 6.3.6 RONDAS DE TRABAJO

El establecimiento de los marcadores requiere habitualmente de tres rondas de trabajo (Lin, 2006). Cada ronda tiene el objetivo de generar un espacio donde puedan debatirse las diferencias y perspectivas de los panelistas, ya sea en los grupos de trabajo pequeños como en el grupo total. La intención del debate y de las rondas es colaborar al consenso y reducir las diferencias entre los panelistas. Ciertas veces, las discusiones no generan ningún cambio en los puntajes de corte establecidos; en otras, los cambios son pequeños pero pueden resultar significativos en la práctica. Además, la discusión en los grupos permite incrementar la seguridad y confianza de los jueces en su tarea y hace que, una vez establecidos los puntos de corte, muestren su acuerdo y acepten de mejor modo los resultados finales (Hambleton, 2001).

**Ronda 1.** Los objetivos principales de la primera ronda son: (a) establecer la ubicación del primer marcador y (b) favorecer el debate en los grupos pequeños.

En la primera ronda, los panelistas recibirán el cuadernillo, el cuadernillo complementario (si corresponde), un formulario en el que completar sus marcas (ver figura 3) y una serie de marcadores. La idea de que los panelistas deban volcar sus resultados sobre el formulario impreso además de colocar los marcadores tiene la intención de generar una segunda instancia donde los jueces presten mayor atención sobre su decisión (Cizek & Bunch, 2007).



**EXAMEN DEL ÁREA:**

**GRADO EVALUADO:**

**NOMBRE DEL PANELISTA:**

**INSTRUCCIÓN:** Por favor, complete los espacios en blanco del siguiente formulario con el número del ítem seleccionado para cada nivel de desempeño. Recuerde que tiene que colocar el número de ítem que se encuentra en la esquina derecha de la hoja del cuadernillo.

**RONDA 1**

| | Bajo/Medio | Medio/Alto |
|---|---|---|
| Nº de Ítem del cuadernillo | | |

**RONDA 2**

| | Bajo/Medio | Medio/Alto |
|---|---|---|
| Nº de Ítem del cuadernillo | | |

**RONDA 3**

| | Bajo/Medio | Medio/Alto |
|---|---|---|
| Nº de Ítem del cuadernillo | | |

**NOTAS:**

**Figura 3.** Ejemplo de formulario de respuesta para los panelistas. **Nota.** Adaptado de Cizek & Bunch (2007), p. 182.

En esta primera ronda los evaluadores trabajarán en grupos de tres a cinco participantes (Cizek & Bunch, 2007). Se discute qué evalúa cada indicador y qué es lo que hace que un ítem sea más difícil que el anterior. El propósito es que los jueces discutan y determinen el contenido que deberían dominar los estudiantes posicionados en cada nivel de desempeño. Cada evaluador coloca de manera independiente su marcador en el ítem que le parece adecuado. Un marcador es situado para cada uno de los puntos de corte seleccionados (e.g. si hay tres niveles de desempeño deberán situar dos marcadores). Los indicadores que estén situados antes del primer marcador deben poder ser respondidos de manera correcta por todos los estudiantes de ese nivel de desempeño con una probabilidad de .67.

**Puntos de corte preliminares.** Al finalizar la primera ronda, los asistentes de trabajo reunirán los formularios completados por los panelistas. Estos datos se ingresarán en una computadora para poder brindar a los panelistas un panorama descriptivo de los resultados en la devolución de la primera ronda. Esta información permitirá a los jueces observar dónde quedaron posicionados sus marcadores y comparar su trabajo con el del resto de los participantes, y dará una cierta idea de dónde se posiciona la media del grupo y un parámetro de rangos mínimo y máximo obtenido para cada marcador.

Cada marcador de cada panelista se corresponderá con un valor de habilidad (theta). Para la obtención de los puntajes de corte se promedian los distintos valores correspondientes a cada uno de los panelistas. Una vez obtenidos estos puntos de corte preliminares se procede a calcular la proporción de estudiantes en cada uno de los tres niveles de desempeño. Todas estas tareas deben ser documentadas en favor de resguardar la claridad del procedimiento.

**Advertencias sobre los puntajes de corte obtenidos con el método Bookmark.** En algunas aplicaciones del método, los puntos de corte han sido obtenidos tomando la media de la página del cuadernillo. Por ejemplo, si la media del grupo se establece en la página 29, ese número hubiese sido tomado como puntaje bruto (e.g. Buckendahl, Smith, Impara & Plake, 2002). La lógica de establecer un punto de corte asociado a la media del valor (theta) identificado por los jueces se debe a que los participantes ubican su marcador en el último ítem del cuadernillo que consideran que un estudiante limítrofe tiene menos de un 2/3 de probabilidades de responder correctamente (Cizek & Bunch, 2007). Eso implica que el individuo límite aún tiene ciertas probabilidades de contestar correctamente e incluso de contestar correctamente algunos de los ítems subsiguientes.
El presupuesto de los modelos basados en la TRI es que cada estudiante tiene una probabilidad de contestar cada ítem de manera correcta y que esta probabilidad puede ser estimada.

**Devoluciones de la primera ronda.** Siguiendo las recomendaciones de Cizek y Buch (2007) en la figura 4 se muestra un ejemplo de la información que se brinda como devolución, luego de la primera ronda de trabajo. Normalmente, la mayor variabilidad de las ubicaciones para los marcadores se da en esta primera ronda (Lewis et al., 1998). En esta etapa, sólo se muestran los números de ítems del cuadernillo donde fueron situados los marcadores. De esta forma los participantes tienen una visión global de cómo ha sido el trabajo del resto de los panelistas. La figura muestra además los espacios vacíos, es decir, donde nadie colocó marcadores para ninguno de los niveles de desempeño. En las rondas siguientes, la atención dirigida a ese espacio disminuirá y cobrará mayor importancia la discusión y consideración de aquel rango de ítems que muestre mayores dificultades.

DISTRIBUCIÓN DE LOS MARCADORES - LENGUA 6° GRADO DE PRIMARIA - TERCERA RONDA

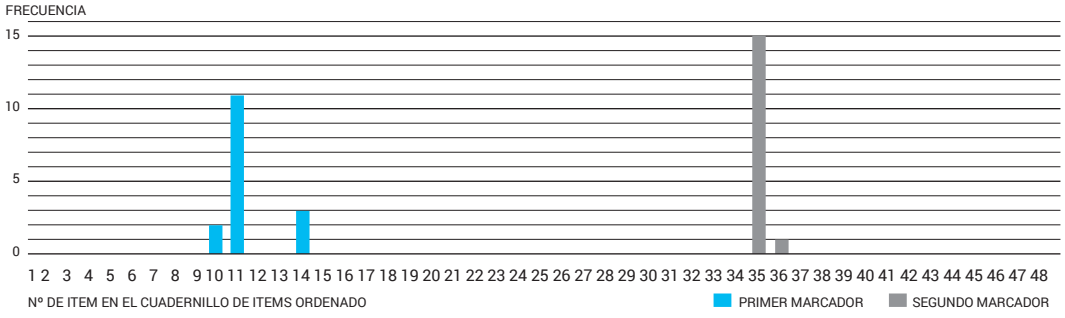


**Figura 4.** Resultados de la tercera ronda de trabajo de Lengua de 6° grado de primaria.
Taller Bookmark, Aprender 2016.

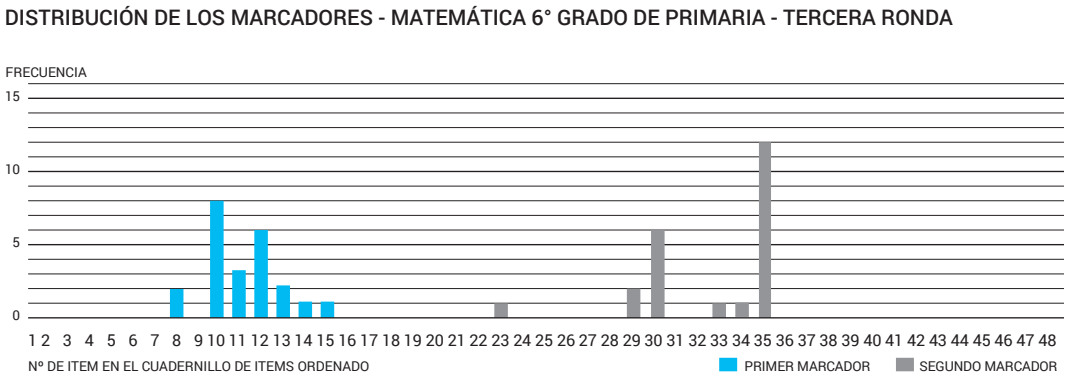


**Figura 5.** Resultados de la tercera ronda de trabajo de Matemática de 6° grado de primaria.
Taller Bookmark, Aprender 2016

En la figura 4, se muestra el continuo de ítems de todo el cuadernillo de Lengua de 6° grado y los distintos marcadores establecidos para cada punto de corte. En este sentido, se evidencia cuáles ítems muestran un mayor acuerdo entre los diferentes jueces y cuáles menor acuerdo. En la figura 4, se puede observar poca dispersión entre los ítems seleccionados para el primer marcador, estando concentradas todas las respuestas entre el ítem 8 y el 15. En lo que refiere al segundo marcador se observa un gran acuerdo en el ítem 35, siendo 15 docentes quienes seleccionaron este ítem y una distribución de las elecciones de los jueces relativamente concentrada en un rango de ítems acotados (del 29 al 35), a excepción de un único participante que optó por elegir el ítem 23.

Por otro lado, en la figura 5 se puede notar que mientras algunos jueces colocaron el segundo marcador (para el punto de corte del nivel Medio-Alto) en el ítem 30, la gran mayoría optó por el ítem 35. En este sentido, el observar el gráfico ayuda a poner de manifiesto las diferencias entre las competencias y conocimientos que unos consideran correspondientes a cada nivel y poder debatir acerca de ellas.

Además de los datos normativos, todos los procedimientos de establecimiento de estándares suelen brindar datos de impacto. Es decir, datos que incluyan cómo afectarían el establecimiento de determinados puntos de corte a la población estudiada (i.e. distribución porcentual de estudiantes según cada nivel de desempeño en función del puntaje de corte estimado). No hay un momento adecuado para la presentación de este material. Puede ser luego de la primera, segunda o tercera ronda. Sin embargo, de acuerdo con Cizek y Buch (2007) es importante poder facilitar este tipo de información inmediatamente después de la primera ronda, ya que cuanto más tarde se presenten estos datos, menor será el impacto que tengan sobre las decisiones de los jueces. El propósito de brindar esta información es dar una idea global del conjunto de juicios provistos por el grupo. Hambleton (2001) señala que no todos los actores involucrados en el establecimiento de estándares suelen estar de acuerdo en mostrar estos datos a los panelistas. Algunos actores políticos prefieren no mostrar estos datos, y llegado el momento, ser ellos quienes modifiquen los puntos de corte obtenidos teniendo en cuenta esta información.

Sintetizando, Jornet Meliá y Backhoff (2006) señalan que en la fase de retroalimentación de cualquiera de las tres etapas se deberá incluir: (a) el grado de congruencia de los jueces para cada ítem seleccionado; (b) el número de reactivos que definan cada nivel de desempeño y las discrepancias en la identificación de reactivos entre los jueces; y (c) la distribución porcentual de individuos en cada nivel de logro.

**Ronda 2.** Las tareas de los participantes en la segunda ronda son básicamente idénticas a las que deben realizar en la primera. Se apunta a que los panelistas reconsideren la posición de sus marcadores teniendo en cuenta los contenidos de los ítems y cuáles de ellos corresponden a cada nivel de desempeño.

En la segunda ronda se vuelven a entregar a los participantes los cuadernillos, el formulario donde deben completar sus marcas, los marcadores y un resumen de los resultados de la primera ronda incluyendo los datos normativos y la información de impacto si la hubiese.

Algunos autores sugieren que esta segunda ronda comience con una discusión acerca de los resultados de la primera dentro del grupo grande (Cizek & Bunch, 2007). Luego de la discusión en el grupo numeroso, se continúa el trabajo en los grupos pequeños y, por último, se trabaja individualmente colocando cada juez sus marcadores. Otros autores recomiendan comenzar el trabajo de la segunda ronda en grupos reducidos de tres a cinco participantes (Lin, 2006).

Los grupos pequeños de la segunda ronda pueden estar conformados de igual manera que en la primera o bien conformarse nuevos grupos reasignando aleatoria o intencionalmente a los participantes. La asignación intencional responderá al objetivo de generar debate entre puntos divergentes y poder llegar a un consenso.

Se sugiere que en cada grupo pequeño, los participantes agreguen a sus cuadernillos otros marcadores que representen los lugares donde los colegas del pequeño grupo han colocado sus marcas (Lin, 2006). Siguiendo esta propuesta, cada participante debería tener en su cuadernillo sus dos marcas más las marcas de sus colegas del grupo reducido de tal modo que pueda observar donde los demás colocaron el punto de corte para cada nivel. Para un grupo de seis panelistas, cada panelista tendrá seis marcadores por punto de corte (los suyos y los de sus colegas).

Ya sea en el grupo numeroso o reducido, la actividad inicial se centrará en los resultados obtenidos en la primera ronda. Los líderes de cada grupo pequeño o los asistentes del grupo grande promoverán la discusión focalizando la atención en los marcados de los extremos, es decir en el primer y último marcador correspondiente a cada punto de corte y en los ítems que muestran desacuerdos y/o dificultades particulares. A veces puede ser útil que cada participante explicite y explique las razones por las que considera que el ítem que ha señalado, debería ser el ítem separador. Al finalizar el debate, cada panelista deberá volver a colocar su marcador de manera independiente para cada uno de los puntos de corte completando las actividades de la segunda ronda.

**Devoluciones de la segunda ronda.** Incluirán los datos normativos y podrán sumar la información de impacto. En algunos casos también se sugiere aportar gráficos que ilustren los cambios en las ubicaciones de los marcadores de una ronda a otra de manera individualizada (Herrera Ortiz et al., 2009). En la figura 6 se muestra un ejemplo.

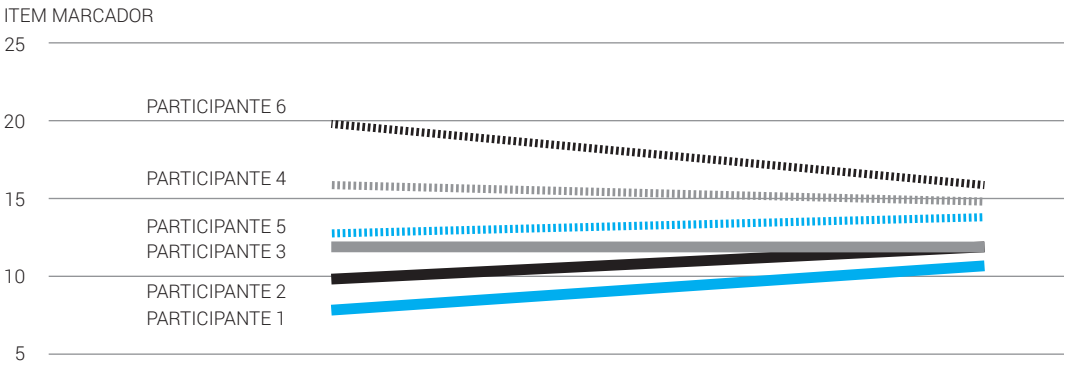


**Figura 6.** Ejemplo de datos mostrados en la devolución de la segunda ronda
**Nota.** Adaptado de Herrera Ortiz et al. (2009), p. 57

**Ronda 3.** Al comenzar la tercera ronda cada participante recibe su cuadernillo con los materiales de la segunda ronda más una síntesis de los resultados de ésta.

En la tercera ronda, los asistentes deberán liderar y conducir la discusión acerca de los datos obtenidos en la ronda anterior. Esta discusión se lleva a cabo en el grupo total. Teniendo en cuenta toda la información recopilada en cada una de las rondas de trabajo se les pide a los participantes que coloquen por tercera y última vez, sus marcadores para cada punto de corte. Se insta a los panelistas a prestar atención en el formulario donde han asentado previamente sus marcas, a examinar esos marcadores y evaluar qué cambios han tenido lugar en las rondas anteriores.

## 6.4 DEFINICIÓN DE LOS PUNTAJES DE CORTE

Tal como se ha mencionado anteriormente, los puntajes de corte son obtenidos mediante la sumatoria y promedio de los valores de habilidad theta ($\theta$), correspondientes a la ubicación de los marcadores para cada nivel de desempeño.

En algunas otras aproximaciones el número de ítem del cuadernillo es tomado directamente como punto de corte. Es decir que, si un juez coloca su marcador en el ítem número 10, el punto de corte recomendado será 10 (Cizek & Bunch, 2007). Esta estrategia ha sido implementada en un estudio realizado por Buckendahl et al. (2002). Otra alternativa se orienta a ordenar los ítems de acuerdo a su valor p, en vez de seguir los valores obtenidos por los modelos TRI. En estos casos, cada ítem del cuadernillo tiene un puntaje bruto determinado, por ejemplo el ítem 10 tendrá un puntaje bruto equivalente a 10 puntos. Básicamente, esta estrategia plantea un modo diferente de escalamiento lo que, según Cizek y Buch (2007), podría tener algunas consecuencias inesperadas y poco controlables.

Jornet Meliá y Backhoff (2006) sugieren otra alternativa para la selección de los puntos de corte. De acuerdo a las pautas propuestas por los autores, los puntos de corte se corresponderán a la mediana del nivel de habilidad correspondiente a los ítems marcados por cada uno de los panelistas. En segundo lugar, señalan algunos criterios de calidad a considerar acabado el proceso. Entre los mismos mencionan: (a) el grado de congruencia entre los jueces para cada punto de corte, tomándose como referencia la distancia entre

los puntos de corte y la desviación en los juicios emitidos; (b) la valoración que hagan los jueces acerca de la representatividad de los puntajes de corte obtenidos y los porcentajes de estudiantes ubicados en cada nivel y; (c) la existencia de una distancia suficiente entre las distintas puntuaciones de corte y, por tanto, entre los niveles de desempeño.

A propósito de estas menciones, nuevos estudios deberían ser realizados para aportar evidencias sobre las ventajas del uso de la media o de la mediana para el cálculo de los puntos de corte. Karantonis y Sireci (2006) señalan que es probable que la mediana disminuya la influencia que los marcadores extremos tienen sobre las puntuaciones de corte, mientras que la media puede resultar una medida más fiel a las variaciones en las respuestas de los jueces.

Una última consideración a tener en cuenta es que los puntajes de corte obtenidos en las rondas de trabajo podrán ser modificados por las autoridades de determinados sectores (e.g. secretarías, ministerios) con el objetivo de poder articular la información obtenida en diferentes grados o grupos de panelistas (Hambleton, 2001). En el caso de ser necesario, esto deberá ser adecuadamente informado a fin de resguardar la validez del procedimiento aplicado.



## 6.5 FINALIZACIÓN DEL PROCEDIMIENTO BOOKMARK: REDACCIÓN DE LOS DESCRIPTORES DE LOS NIVELES DE DESEMPEÑO

En el procedimiento Bookmark, los descriptores suelen redactarse o refinarse, según corresponda, en la etapa final del trabajo. Una vez obtenidos los puntajes de corte finales se deberán dejar por escrito los descriptores para cada nivel de desempeño. Éstos describen el conocimiento específico, las habilidades y las capacidades de los estudiantes esperables para cada uno de los niveles. Se espera que reflejen y sinteticen el contenido de los ítems incluidos en cada nivel.

Estos descriptores deben ser útiles para poder identificar qué debe conocer y ser capaz de hacer un estudiante de cada nivel. Los indicadores que se encuentran ubicados antes del marcador del primer puntaje de corte reflejan un contenido que se espera que los estudiantes dominen y puedan responder correctamente al menos con un 67% de probabilidades. De acuerdo a esta perspectiva, los descriptores representan una extensión de la tarea correspondiente al establecimiento de los puntos de corte.

Los descriptores podrán ser redactados por los panelistas, por el equipo técnico o grupo de especialistas en el área de conocimiento evaluado, o por ambos como una tarea conjunta (Cizek et al., 2004; Herrera Ortiz et al., 2009; Jornet Meliá & Backhoff, 2006, 2008; Perie, 2008). Por otra parte, se espera que los descriptores sean graduales, es decir, que acompañen las diferencias cualitativas entre un nivel de desempeño y otro. Para ello es recomendable que quienes redacten los descriptores conozcan cuáles son los contenidos estándares sobre los que se apoya cada evaluación.

Jornet Meliá y Backhoff (2006) ofrecen una lista de verbos que pueden ayudar a orientar a quienes realicen la tarea a identificar cuatro diferentes niveles de logro y su graduación: (1) reconocer, encontrar, identificar, nombrar, señalar, elegir; (2) comprender, agrupar, asociar, organizar, clasificar, jerarquizar, interpretar; (3) utilizar, anticipar, predecir, parafrasear, reconstruir, interpretar, resumir, explicar, integrar, solucionar, cambiar; (4) aplicar, argumentar, criticar, cuestionar, opinar, reflexionar, valorar, convertir, demostrar, extra-

polar, planear, transformar. Esta lista es sólo orientativa y deberá ser adaptada para aquellos casos en donde los niveles de desempeño sean menores (e.g. tres niveles de logro). Independientemente de quiénes integren los grupos para el desarrollo de los descriptores se sugiere que el trabajo sea realizado en grupos pequeños donde se debata brevemente cuáles son los conocimientos y habilidades de cada caso y luego se realice una puesta en común (Jornet Meliá & Backhoff, 2006).

De acuerdo con Herrera Ortiz et al. (2009) los descriptores deberían mostrar qué es lo que los estudiantes de cada nivel pueden hacer en lugar de los contenidos o temas evaluados en las pruebas. Los autores recomiendan, además, complementar esta información con datos y ejemplos que expliciten los conocimientos y habilidades que manifiestan los estudiantes de cada nivel. Otros trabajos proponen seleccionar e incluir un ítem representativo de cada disciplina, grado y nivel de desempeño como muestra del tipo de actividades que se espera de los estudiantes pertenecientes a dicho nivel (Jornet & Backhoff, 2006)[5]. Por otro lado, se aconseja que toda la información relativa a los descriptores de cada nivel y sus ejemplos sea desarrollada en un párrafo que sintetice de manera global las características principales.

Jornet Meliá y Backhoff (2006) sugieren que el lenguaje utilizado en la elaboración de los descriptores sea: (a) técnicamente correcto y preciso; (b) comprensible para la mayoría de los individuos que conforman la sociedad (no únicamente técnicos y especialistas); (c) incluya las competencias, habilidades, destrezas, conocimientos y logros y no haga referencia a contenidos específicos; (d) contemple el uso de términos que identifiquen niveles diferenciales en cada una de las competencias que se señalen (para ello pueden utilizarse la clasificación de verbos sugerida con anterioridad); (e) evite el uso de un lenguaje ambiguo que pueda dar lugar a una mala interpretación, y siempre que sea necesario, aclarar; y (f) no incluya términos que puedan denotar discriminación y en su lugar promover un uso del lenguaje respetuoso.

Una vez finalizada la redacción de los descriptores, se procederá a recolectar las últimas evidencias acerca de la validez y confiabilidad de la metodología utilizada. Como ya se ha recomendado, se espera que durante todo el proceso de establecimiento de estándares se lleve un cuidadoso registro y documentación de cada una de las etapas. En esta última oportunidad, se les solicitará a los jueces que contesten una breve encuesta acerca de cómo les ha resultado el proceso. Existen diferentes escalas para evaluar la percepción final de los jueces acerca de las tareas realizadas. Algunos autores como Hambleton (2001) proponen una serie de escalas que abordan diferentes cuestiones en mayor detalle. Sin embargo, dado que el proceso de establecimiento de puntos de corte es extenso y los jueces suelen estar fatigados hacia el final de las tareas, otros autores proponen utilizar una única escala breve que sintetice los aspectos fundamentales (Cizek et al., 2004; Herrera Ortiz et al., 2009). En la figura 7 se muestra la escala presentada por Cizek et al. (2004), traducida al español.

5 En estos casos, se recomienda un especial cuidado en la selección de los ítems ya que dichos indicadores luego se publicarán como ejemplos en diferentes informes y/o comunicaciones de la institución que lleve a cabo la evaluación.

A continuación se presenta una serie de afirmaciones respecto de las tareas que ha realizado en los últimos tres días. Por favor, lea cada frase y señale si está "De acuerdo" o "En desacuerdo". Al terminar puede agregar aquellos comentarios que considere podrían enriquecer este proceso a futuro.

| | | DE ACUERDO | EN DESACUERDO |
|---|---|---|---|
| | Durante la capacitación pude entender claramente el propósito de las tareas a realizar | | |
| | Los asistentes y coordinadores de las jornadas explicaron claramente las actividades que debíamos realizar | | |
| | La capacitación y la ronda práctica me ayudó a entender cómo realizar las tareas de cada ronda | | |
| | Revisar los ítems del cuadernillo me ayudó a comprender mejor la evaluación | | |
| | Las etiquetas de los niveles me parecieron claras y útiles | | |
| ○ | Las discusiones en los grupos pequeños y en el grupo completo me ayudaron a comprender el procedimiento y realizar la tarea con mayor eficacia | | |
| | El tiempo previsto para las rondas de discusiones fue adecuado | | |
| ○ | Todos los participantes del pequeño grupo en el que participaba tuvieron las mismas oportunidades para expresar sus opiniones y puntos de vista | | |
| | Fui capaz de seguir las instrucciones y completar el formulario en cada ronda | | |
| | Las discusiones realizadas después de la primera ronda me resultaron útiles | | |
| | Las discusiones realizadas después de la segunda ronda me resultaron útiles | | |
| | Los datos brindados en las devoluciones de cada ronda por los asistentes y coordinadores me resultaron útiles para realizar la tarea | | |
| | Me siento seguro de los marcadores establecidos y creo que son apropiados para la población implicada | | |
| | Las facilidades de alojamiento, viáticos y servicios brindados por los coordinadores generaron un buen clima de trabajo | | |
| | Comentarios/Sugerencias: | | |

**Figura 7.** Ejemplo de escala autoadministrable para la evaluación de la participación de los jueces.
**Nota.** Adaptado de Cizek et al 2004 p.45

# 7. MÉTODO BOOKMARK EN APRENDER 2016



Para facilitar el trabajo del Taller Bookmark, Aprender optó por mantener la misma cantidad de niveles de desempeño y las mismas etiquetas utilizadas en el Operativo Nacional de Evaluación (ONE 2010 y 2013). Por lo cual los niveles de desempeño fueron tres y las etiquetas de los niveles de desempeño se correspondieron con Bajo, Medio y Alto.

En Aprender 2016, la tarea de definir los marcadores se le encomendó a un grupo conformado por 201 docentes en ejercicio de las 24 jurisdicciones del país, pertenecientes a escuelas estatales y privadas, de ambos géneros, de diferentes edades y años de experiencia en el cargo[6].

En la semana del 13 al 17 de febrero del 2017, se llevaron a cabo los talleres de trabajo Bookmark. Apoyados en la información correspondiente a un subconjunto de los ítems de cada prueba, ordenados según su nivel de dificultad, los grupos debieron examinarlos cuidadosamente para poder definir los marcadores. Se analizó la dificultad de los ítems, las operaciones cognitivas que entran en juego para contestarlas de forma correcta, las alternativas de respuesta que pueden facilitar o complejizar la elección de la respuesta correcta, entre otras cuestiones.

El debate se llevó a cabo durante tres rondas sucesivas. Para cada una los docentes debieron completar un formulario donde consignaron de manera individual los ítems elegidos como marcadores. Esa información se utilizó luego para abrir el debate en la ronda siguiente y, al final, se eligieron como marcadores aquellos consignados por los docentes luego de la tercera ronda. El valor theta promedio de los ítems elegidos como marcadores en la última ronda es lo que define los puntos de corte de cada nivel.

## 7.1 DEFINICIÓN DE LOS PUNTOS DE CORTE EN APRENDER 2016

Una vez obtenidos los resultados de cada ronda, insumo principal utilizado para el primer paso en el establecimiento de los puntos de corte, se realizó un análisis de cada uno de los grupos por separado para evaluar los resultados. En casi la totalidad de los casos, se encontró una concentración de las respuestas de los participantes en un rango menor a 12 puntos.

En Ciencias Sociales de 5°/6° año y en el segundo marcador de Matemática de 5°/6° año y de Lengua de 3° grado, no se identificó un consenso claro entre los diferentes participantes. Básicamente, el rango de los marcadores fue mayor a 13 puntos en todos los casos mencionados además de mostrar una mayor dispersión en las frecuencias. En función de estos datos, se decidió convocar a tres expertos de cada una de estas disciplinas y años para poder, a partir de los datos de los docentes, tomar una decisión respecto de cuál era el marcador apropiado para cada caso.

6 Como Anexo de este documento se encuentra la lista de docentes participantes.

En esta convocatoria prestaron su colaboración personal experto del equipo pedagógico de la Secretaría de Evaluación Educativa y de la Secretaría de Innovación y Calidad Educativa (Gestión Curricular e INFOD). En dichas reuniones se explicó a los expertos en qué consistía el establecimiento de los puntos de corte, cuál había sido la tarea de los docentes participantes en los talleres Bookmark, se les mostraron los resultados encontrados en la tercera ronda de cada una de las áreas y años que presentaban dificultades y se les solicitó que, teniendo en consideración estos datos, pudiesen emitir un juicio experto fundamentando su elección.

Una vez que se expusieron e intercambiaron opiniones y argumentos, se consensuó la elección de un solo ítem para el primer y segundo marcador de Ciencias Sociales 5°/6° año y para el segundo marcador de Matemática 5°/6° año y Lengua 3° grado.

## 7.2 DEFINICIÓN DEL CUARTO NIVEL DE DESEMPEÑO

Con posterioridad a la definición de los puntos de corte, se tomó la decisión de cambiar la denominación de los niveles de desempeño (Alto, Medio y Bajo) que pasaron a denominarse: Básico, Satisfactorio y Avanzado. Esto no implicó cambios en los marcadores que se habían fijado. Los puntos de corte elaborados en el marco de los talleres Bookmark determinaron los niveles de desempeño de los resultados de las pruebas Aprender 2016, aplicándose únicamente un cambio en las etiquetas de dichos niveles.

Se realizó luego una modificación adicional: el nivel Básico (antes denominado Bajo) se subdividió en dos para responder a la necesidad de agudizar el análisis de este grupo. Quedó definido entonces, por un lado, el nivel denominado Por debajo del nivel básico que incluye a los estudiantes cuyo puntaje se distancia en más de un 25% respecto del punto de corte del nivel Satisfactorio. Por el otro lado, se mantuvo la denominación Básico pero incluyendo ahora a los estudiantes cuyos puntajes no superan esa diferencia. Esta subdivisión y el criterio utilizado es similar a los ajustes que suelen realizarse en las evaluaciones estandarizadas a nivel regional e internacional.

De esta forma, quedaron definidas cuatro categorías:

| |
|---|
| **POR DEBAJO DEL NIVEL BÁSICO** |
| **BÁSICO** |
| **SATISFACTORIO** |
| **AVANZADO** |

Para cada uno de los cuatro niveles de desempeño se elaboraron descriptores de nivel los que refieren con detalle a los saberes y capacidades que deben poseer los estudiantes para poder alcanzar cada uno de los niveles de desempeño. Los contenidos de aprendizajes detallados para cada descriptor de nivel siguen los lineamientos definidos en los Núcleos de Aprendizaje (NAP) acordados a nivel nacional por las 23 provincias y la Ciudad Autónoma de Buenos Aires en el marco del Consejo Federal de Educación.

# 8. CONCLUSIONES Y RECOMENDACIONES FINALES



El establecimiento de estándares es un proceso complejo que implica diferentes etapas e involucra a una pluralidad de actores y disciplinas. Las consecuencias de los procedimientos para el establecimiento de estándares pueden afectar a determinados sectores de la sociedad, o en los casos donde se refieren a evaluaciones educativas de nivel nacional y/o internacional, a la sociedad en su conjunto.

El método Bookmark es uno de los más utilizados en la actualidad para la determinación de los puntos de corte en las evaluaciones del ámbito educativo. El método muestra diferentes ventajas que facilitan el trabajo de los panelistas y disminuye el grado de error en los resultados. Ahora bien, tal como se ha mencionado anteriormente, la validez y confiabilidad del método no puede por sí misma ser suficiente para garantizar el éxito de los procedimientos para el establecimiento de estándares. En este punto también se hace preciso mencionar que los puntos de corte se encuentran irremediablemente ligados al diseño y nivel de dificultad de las pruebas administradas. Si la prueba es sencilla, probablemente el primer punto de corte tenga por debajo una cantidad numerosa de ítems. Esto no es una dificultad del método Bookmark sino una dificultad de cualquier método para el establecimiento de estándares. Es decir, la validez de la prueba debe ser cuidadosamente evaluada antes de la aplicación de los métodos para la determinación de los puntos de corte.

Para finalizar y a pesar de todas las fortalezas que han sido referidas al método Bookmark, se han de señalar también algunas limitaciones. De acuerdo con Lin (2006), algunas de éstas son: (a) la elección de un grado de probabilidad de respuesta (.50 o .67); (b) el desacuerdo que muestran algunos jueces respecto del ordenamiento de determinados ítems en el cuadernillo; (c) la incapacidad para tener en cuenta otros parámetros además del nivel de dificultad o del nivel de habilidad requerido (e.g. no es posible ponderar los contenidos, tal vez en Matemática es más importante que un estudiante pueda resolver un determinado problema que demuestre saber un determinado contenido) y; (d) algunas restricciones propias de los modelos TRI. A pesar de estas limitaciones y como ya se ha enunciado en otros apartados, las ventajas del método Bookmark son amplias respecto de otros métodos para el establecimiento de puntos de corte basados en el juicio de experto.

Por un lado, el establecimiento de estándares tiene por finalidad generar evaluaciones confiables y válidas que permitan conocer los aprendizajes y habilidades que dominan los sustentantes de diferentes niveles de desempeño. Por otro lado, se busca favorecer (mediante este conocimiento y los datos obtenidos) el diseño y la implementación de políticas educativas a fin de que determinados aprendizajes comunes a cada área disciplinar y a cada grado estén garantizados, promoviendo una mejora de la calidad educativa de la población en su conjunto. Se espera con este trabajo contribuir a tal fin.

# 9. REFERENCIAS

Abbott, M. L. (2006). Setting cut-scores for complex performance assessments: A critical examination of the analytic judgment method. Alberta Journal of Educational Research, 52(1), 25-35.

American Educational Research Association, American Psychological Association, National Council on Measurement in Education (1999). Standards for educational and psychological testing. Washington, DC: American Psychological Association.

Angoff, W. H. (1971). Scales, norms, and equivalent scores. In R. L. Thorndike (Ed.), Educational measurement (pp. 508-600). Washington, DC: American Council on Education.

Beck, M. (2003, April).Standard setting: If it is science, it's sociology and linguistics, not psychometrics. Paper presented at the annual meeting of the National Council on Measurement in Education, Chicago, IL

Beck, A. T. & Steer, R. A. (1987). Manual for the revised Beck Depression Inventory. San Antonio, Tex: Psychological Corporation.

Bejar, I. (1983). Subject matter experts' assessment of item statistics. Applied Psychological Measurement, 7,303-310.

Bender, L. A. (1938). A visual motor Gestalt test and its clinical use. American Orthopsychiatric Association Research Monograph 3. New York: American Orthopsychiatric Association.

Berk, R.A. (1986). A consumer's guide to setting performance standards on criterion-referenced tests. Review of Educational Research, 56, 137-172.

Beuk, C.H. (1984). A method for reaching a compromise between absolute and relative standards in examinations. Journal of Educational Measurement, 21, 147-152.

Brown, W.J. (2001). Social, educational, and political complexities of standards setting. En G.J. Cizek (ed.). Standard setting performance standards: Concept, methods and perspectives (373-386), Mahwah, N.J: Erlbaum

Buckendahl, C. W., Smith, R. W., Impara, J. C.; & Plake, B. S. (2000, October). A comparison of the Angoff and Bookmark Standard Setting Methods. Paper presented at the Annual Meeting of the Mid-Western Educational Rsearch Association, Chicago.

Cizek, G. J. (1993). Reconsidering standards and criteria. Journal of Educational Measurement, 30(2), 93-106.

Cizek, G. J. (1996). Standard setting guidelines. Educational Measurement: Issues and Practice, 15(1), 13-21.

Cizek, G. J. (2001). Conjectures on the rise and call of standard setting: An introduction to context and practice. In G. J. Cizek (Ed.), Setting performance standards: Concepts, methods, and perspectives (pp. 3-17). Mahwah, NJ: Erlbaum.

Cizek, G.J. (2006). Standard setting. En S.M. Doning y T.M. Haladyna (eds.). Handbook of Test Development (225-258), New Jersey: Lawrence Erlbaum Associates, Inc.

Cizek, G. J. & Bunch, M. B. (2007). The bookmark method. In Standard setting (pp. 154-191). Thousand Oaks, CA: Sage.

Cizek, G. J., Bunch, M. B. & Koons, H. (2004). Setting performance standards: Contemporary methods. Educational Measurement, 23(4), 31–49.

Comisión para el Desarrollo y Uso del Sistema de Medición de la Calidad de la Educación - Unidad de Curriculum y Evaluación (2006). Resumen Ejecutivo. Aplicación de la metodología para establecer Puntajes de Corte en las pruebas SIMCE 4° Básico de: Lectura, Matemática, Compresión del Medio Natural, Compresión del Medio Social y Cultural. Gobierno de Chile. Obtenido el 13 de octubre de 2016 en: http://www.agenciaeducacion.cl/wp-content/uploads/2013/02/Metodologia_Puntajes_de_Corte_Pruebas_SIMCE_4_B_2008.pdf

Cooper-Loomis, S., & Bourque, M. L. (2001). From tradition to innovation: Standard setting on the National Assessment of Educational Progress. In G. Cizek (Ed.), Setting performance standards: Concepts, methods, and perspectives (pp. 175-217). Mahwah, NJ: Erlbaum.

Cortada de Kohan, N., Macbeth, G., & López Alonso, A. (2008). Técnicas de Investigación Científica. Buenos Aires, Argentina: Lugar.

Cruz Ampuero, G., Espinoza Pezzia, G., Montané Lores, A. &, Rodríguez Cuellar, C. (2001). Informe Técnico de la Consulta sobre Puntos de Corte para la Evaluación Nacional 2001. Ministerio de Educación. Unidad de Medición de la Calidad Educativa. Lima, Perú. Obtenido el 13 de octubre de 2016 en: http://www2.minedu.gob.pe/umc/admin/images/menanexos/menanexos_45.pdf

De Gruijter, D.N. (1985): Compromise methods for establishing examination standards, Journal of Educational Measurement, 22, 263-269.

Department of Public Instruction & CTB/McGraw-Hill (2003). 2003 Standard Setting or Wisconsin Knowledge and Concepts Examination (WKCE). Obtenido el 18 de octubre de 2016 en: https://dpi.wi.gov/sites/default/files/imce/assessment/pdf/2003%20Standard%20Setting%20for%20the%20WKCE.pdf

Ebel, R. L. (1962). Content standard test scores. Educational and Psychological Measurement, 22, 15-25.

Egan, K., Barton, K., & Roeber, E. (2015). ISTEP+Standard Setting. Obtenido el 18 de octubre en: http://www.in.gov/sboe/files/Indiana_Istep_Standard_Setting_Memo_final.pdf

García, P. E., Abad, F. J., Olea, J., & Aguado, D. (2013). A new IRT-based standard setting method: Application to eCat-Listening. Psicothema, 25(2), 238-244.

Gigerenzer, G. & Hoffrage, U. (1995). How to improve Bayesian reasoning without instruction: Frequency formats. Psychological Review, 102, 684-704.

Glaser, R. (1963). Instructional technology and the measurement of learning out-comes: Some questions. American Psychologist, 18, 519-521.

Glass, G. V. (1978). Standards and criteria. Journal of Educational Measurement, 15, 237-261.

Hambleton, R. H. (2001). Setting performance standards on educational assessments and criteria for evaluating the process. In G. J. Cizek (Ed.), Setting performance standards: Concepts, methods, and perspectives (pp. 89–116). USA: Taylor & Francis.

Hambleton, R. K., Jaeger, R. M., Plake, B. S. & Mills, C. N. (2000). Handbook for setting standards on performance assessment. Washington, DC: Council of Chief State School Officers.

Hambleton, R. K., & Plake, B. S. (1995). Using an extended Angoff procedure to set standards on complex performance assessments. Applied Measurement in Education, 8, 41-56.

Hambleton, R.K. & Swaminathan, H. (1978). Criterion-Referenced Testing and Measurement: A review of technical issues and developments. Review of Educational Research, 40, 1-47.

Hambleton, R. K., Swaminathan, H., & Rogers, J. (1991). Fundamentals of item response theory. Beverly Hills, CA: Sage.

Herrera Ortiz, M., Benavides Posadas, D. &, Monroy Cazorla, L. (2009). Establecimiento de estándares en un examen criterial. Cuaderno Técnico 3. Méjico, D.F.: Centro Nacional de Evaluación para la Educación Superior.

Hofstee, W.K.B. (1983): The case for compromise in educational selection and grading. In S.B: Andersony J.S. Helmick (Eds.), On educational testing (pp. 109-127). San Francisco, CA: Jossey-Bass.

Huynh, H. (2000, April). On item mappings and statistical rules for selecting binary items for criterion referenced interpretation and bookmark standards settings, Paper presented at the annual meeting for the National Council for Measurement in Education, New Orleans.

Ical Choc, P., Xol Choc, S., Ajú, M., Pocon, A. I., Magzu, J., García, R. et al. (2009). Informe de taller. Uso de la metodología "Separador/Bookmark"; basado en resultados de docentes bilingües en cuatro idiomas mayas. Programa Estándares e Investigación Educativa. USAID, Guatemala. Obtenido el 13 de octubre de 2016 en: http://pdf.usaid.gov/pdf_docs/Pnadr972.pdf

Impara, J. C., Giraud, G., & Plake, B. S. (2000, April). The Influence of Providing Target Group Descriptors When Setting a Passing Score. Paper presented at the Annual Meeting of the American Educational Research Association. New Orleans, LA.

Impara, J.C. & Plake, B.S. (1997). Standard setting: An alternative approach. Journal of Educational Measurement, 34, 353-366.

Jaeger, R. M. (1978). A proposal for setting a standard on The North Caroline High School. Paper presented at the spring meeting of the North Caroline Association for Research in Education, Chapell Hill.

Jaeger, R. M. (1995). Setting performance standards through two-stage judgmental policy capturing. Applied Measurement in Education, 8, 15–40.

Jornet Meliá, J. M., & Backhoff, E. E. (2006). Manual Técnico. Establecimiento de niveles de competencia. Dirección de Pruebas y Medición. México: Instituto Nacional para la Evaluación de la Educación.

Jornet Meliá, J. M., & Backhoff, E. E. (2008). Modelo para la determinación de niveles de logro y puntos de corte de los EXCALE. Cuadernos de Investigación. México: Instituto Nacional para la Evaluación de la Educación.

Jornet Meliá, J. M. & González Such, J. (2009). Evaluación criterial: determinación de estándares de interpretación (EE) para pruebas de rendimiento educativo. Estudios sobre Educación, 16, 103-123.

Kahl, S. R., Crockett, T. J., & DePascale, C. A. (1994, June). Using actual student work to determine cut-scores for proficiency levels: New methods for new tests. Paper presented at the National Conference on Large-Scale Assessment, Albuquerque.

Karantonis, A., & Sireci, S. G. (2006). The bookmark standard setting method: A literature review. Educational Measurement: Issues and Practice, 25(1), 4-12.

Kingston, N.M., Kahl, S.R., Sweeney, K.P., & Bay, L. (2001). Setting performance standards using the Body of Work method. In G.J. Cizek (Ed.), Setting Performance Standards: Concepts, Methods and Perspectives (pp. 219-247). Mahwah, NJ: Erlbaum.

Lewis, D. M., & Green, D. R. (1997, June). The validity of performance level descriptors. Paper presented at the Council of Chief State School Officers National Conference on Largescale Assessment, Phoenix, AZ.

Lewis, D. M., Mitzel, H. C., & Green, D. R. (1996, June). Standard setting: A bookmark approach. In D. R. Green (Chair), IRT-based standard setting procedures utilizing behavioral anchoring. Symposium conducted at the Council of Chief State School Officers National Conference on Large-Scale Assessment, Phoenix, AZ.

Lewis, D. M., Green, D. R., Mitzel, H. C., Baum, K., & Patz, R. J. (1998, June). The bookmark standard setting procedure: Methodology and Recent Implementations. Paper presented at the National Council for Measurement in Education annual meeting, San Diego, CA.

Lewis, D. M., Mitzel, H. C., Green, D. R. & Patz, R. J. (1999). The bookmark standard setting procedure. Monterey, CA: McGraw-Hill.

Lewis, D. M., Mitzel, H. C., Mercado, R. L., & Schulz, M. (2012). The Bookmark Standard Setting Procedure. In G. J. Cizek (Ed.), Setting performance standards: Foundations, methods and innovations (pp. 225-253). New York, NY: Routledge.

Leyva Barajas, Y. E. (2011). Una reseña sobre la validez de constructo de pruebas referidas a criterio. Perfiles Educativos, 33(131), 131-154.

Lin, J. (2006). The Bookmark Procedure for Setting Cut-Scores and Finalizing Performance Standards: Strengths and Weaknesses. Alberta Journal of Educational Research, 52(1), 36-52.

Linn, R.L. (1994, October). The likely impact of performance standards as a function of uses: From rhetoric to sanctions. Paper presented at the National Center for Education Statistics and National Assessment Governing Board Joint conference on Standard Setting for Large Scale Assessments, Washington, DC.

Livingston, A., & Zieky, M. J. (1982). Passing scores: A manual for setting standards of performance on educational and occupational tests. Princeton, NJ: Educational Testing Service.

Lord, F. M., & Novick, M. R. (1968). Statistical theories of mental test scores. New York: Addison-Wesley.

Meskauskas, J. A. (1976). Evaluation models for criterion-referenced testing: Views regarding mastery and standard-setting. Review of Educational Research, 46(1), 133-158.

Mitzel, H. C., Lewis, D. M., Patz, R. J., & Green, D. R. (2001). The bookmark procedure: Psychological perspectives. In G. Cizek (Ed.), Setting performance standards: Concepts, methods, and perspectives (pp. 249-281). Mahwah, NJ: Erlbaurm.

Muñiz, J. (2010). Las teorías de los tests: teoría clásica y teoría de respuesta a los ítems. Papeles del Psicológico, 3(1), 57-66.

Nassif, P.M. (1978). Standard-setting for criterion-referenced teacher licensing tests. Paper presented at the annual meeting of the National Council on Measurement in Education, Toronto.

Nellhaus, J. M. (2000). States with NAEPLike Performance Standards. In M. L. Bourque & S. Byrd (Eds.) Student performance standards on the National Educational Assessment of Educational Progress: Affirmation and improvements (pp. 99–130). Washington, DC:

National Assessment Governing Board. Obtenido el 22 de noviembre en: http://files.eric.ed.gov/fulltext/ED450144.pdf

Perie, M. & Smith, S. (2015). Cut Scores for the Kansas Assessment Program. Kansas state department of Education. Obtenido el 18 de octubre en: http://textlab.io/doc/2151988/dr.-marianne-perie--cete-dr.-scott-smith--ksde-presentati...

Perie, M. (2008). A Guide to Understanding and Developing Performance Level Descriptors. Educational Measurement: Issues and Practice, 27(4), 15-29.

Plake, B. S., Hambleton, R. K., & Jaeger, R. M. (1997). A new standard-setting method for performance assessments: The dominant profile judgment method and some field-test results. Educational and Psychological Measurement, 57, 400-411.

Plake, B., & Hambleton, R. (1998, April). A standard-setting method designed for complex performance assessments with multiple performance categories: Categorical assignments of student work. Paper presented at the annual meeting of the American Educational Research Association, San Diego, CA. (ERIC Document Reproduction Service No. ED 422 371)

Plake, B., & Hambleton, R. (2001). The analytic judgment method for setting standards on complex performance assessments. In G. Cizek (Ed.), Setting performance standards: Concepts, methods, and perspectives (pp. 283-312). Mahwah, NJ: Erlbaum.

Popham, W. J. (1978). Criterion-referenced measurement. Eaglewood Cliffs, NJ: Prentice-Hall.

Popham, W.J. & Husek, T.R. (1969). Implication of Criterion-Referenced Test. Applied Psychology Measurement, 4, 469-492.

Radwan, N., & Rogers, W. T. (2006). A critical analysis of the Body of Work Method for setting cut-scores. Alberta Journal of Educational Research, 52(1), 65-79.

Rasch, G. (1960). Probabilistic models for some intelligence and attainment tests. Copenhague: The Danish Institute for Educational Research.

Ricker, K. L. (2006). Setting cut-scores: A critical review of the Angoff and modified Angoff methods. Alberta Journal of Educational Research, 52(1), 53-64.

Shepard, L. A. (1980). Standard setting issues and methods. Applied Psychological Measurement, 4, 447-467.

Shepard, L. A. (1984). Setting performance standards. En R. A. Berk. (Ed.), A guide to criterion-referenced test construction (pp. 169-198.). Baltimore: Johns Hopkins University Press.

Shepard, L., Glaser, R., & Bohrnstedt, G. (Eds.). (1993). Setting performance standards for student achievement.Stanford, CA: National Academy of Education.

Skaggs, G., & Tessema, A. (April, 2001). Item Disordinality with the Bookmark Standard Setting Procedure. Paper presented at the Annual Meeting of the National Council on Measurement in Education, Seattle, WA. Obtenido el 18 de octubre en: http://files.eric.ed.gov/fulltext/ED453275.pdf

Wechsler, D. (1949). Manual for the Wechsler Intelligence Scale for Children. New York: Psychological Corporation.

Zieky, M. J. (1995). Historical perspective on standard-setting for large-scale assessments. In National Assessment Governing Board/National Center for Education Statistics, Joint conference on standard setting for large scale assessment. Vol 2. Proceedings (pp. 1-38). Washington, DC: Government Printing Office.

Zieky, M. J. (2001). So much has changed: How the setting of cutscores has evolved since the 1980s. En G. J. Cizek (Ed.), Setting performance standards: Concepts, methods, and perspectives(pp. 19-52). Mahwah, NJ: Erlbaum.

Zieky, M., Perie, M., & Livingston, S. (2006). A primer on setting cut scores on tests of educational achievement. Princeton, NJ: Educational Testing Service.

# DOCENTES QUE PARTICIPARON DEL TALLER BOOKMARK

Vanesa Anahí Acosta

Daniela Paola Adanto

María Soledad
Del Valle Aguirre

Claudia Margarita
Altamirano

Víctor Modesto Alvares

Mariela María de los
Ángeles Álvarez Palacios

Sonia Andrea Alzugaray

Azucena Amaya

Carolina Natalia Antunes

Ana Apertile

Natalia Soledad Aranzazu

Yamila Gissell Arce

Juan Ramón Arce

Elena Arévalo

María Del Carmen Argüello

Inés Rosana Arredondo

Maria Guadalupe Bach

Sandra Haydee Baigorria

Gonzalo Javier
Baigorria Ocampo

Liliana Baliani

Guillermo Eduardo
Banchieri

Maria Del Luján Baron

Romina Barragan

Walter Damián Barreto

Alejandra Barrientos

Nadia Jaqueline Blanco

Rosa Adelaida Bobis

Georgina Bonventre

Alejandra Bordenave

Karina Briatore

Manuel Alberto Brito

Laura Elizabeth Bueno

Mario Ariel Buschiazzo

Mabel Liliana
del Carmen Cabezas

Gloria Del Valle Canavidez

Patricia Analía Cano

Dino Edmundo Cantero

Myrian Mabel Cantero

Graciela Beatriz
Caradonna Contrera

Silvia Carmona

Joregelina Del Valle Carrizo

María Isabel Carro

Maribel Del Rosario Casa

Oscar Ramiro Castillo

Mariana Inés Catoggio

Gabriela Cazzulo

Marcos Leandro Chinen

Carla Chiolerio

María de los Ángeles
Chirre Navarro

Sol Chocobar

Lucíla Cocco

Dora Beatriz Colman

Marta Conte

Angelina Del Rosario
Contreras

Elba Ivana Correa

Stella Maris Cristeche

Jose Rafael Crocco

Romina D' Emilio

Graciela Noemi Da Silva

Roxana De La Vega Pérez

Graciela De Robbio

Claudia Del Valle Rodríguez

Andrea Daniela Dhó

María Isabel Diaz

Eduardo Ariel Eberle

Guadalupe Elias

Blanca Erazo

Nancy Escobar

Mirian Escobar

Flavia Marcela
Escudero Gordillo

Lucrecia Faccini

Daniel Alejandro Falleau

María Mónica Farfán

Zulma Irene Fernández

Raul Fernandez.

Luciana Ferratto

Rosana Freddi

Erica Alexia Frias

Andrés Galdeano

Romina Valeria Garcia

Noemí Giampietro

Gabriela Giorgetta

María José Giron

Estela Margarita Glionna

Analia Godoy

Sabrina Isabel Gomez

Sonia Erika Gómez

Silvina Soledad Gonzalez

Fernanda Gonzalez

María del Carmen González

Myriam Graciela Gramajo

Sonia Mabel Guerrero

Carmen Gutierrez

Nancy del Carmen
Gutierrez

Gabriela Leticia Gutierrez

Patricia Noemi Herrera

Sandra Herrera

Daiana Hurtado

María Rosa Jauregui

Franco Efrain Jimenez
Silvana Juarez
Silvina Juliá
Noemí Leiva
Marie Line Nicole Lemoigne
Daniel Leoni
Nancy Liendo
Silvina Del Luján López
Gabriela Alejandra Luna
Serena Machaca
María Del Carmen Maciel
Javier Abdón Maidana Rodriguez
Gisela Soledad Mana
David Orlando Mancuso
Rita Martinez
Graciela Martinez
Elizabet Martínez
Ana Elizabeth Mascheroni
Mariana Soledad Maza
Evita Ana Mendez
Osvaldo Adrián Mendez
Liliana Beatriz Merenda
Marisa Guadalupe Micheletti
Adriana Del Valle Migliavacca
Luis Marcelo Ramón Mijalenko
Gabriela Miranda
Paulina Morello
Yolanda Adriana Moya
Miguel Agustín Mullen
Anahí Carla Natali
Martin Naumann
Marilina Navarro Tenca
Walter Daniel Neirot
Felix Paulo Nuñez
Viviana Lorena Odetti
Walter Ceferino Ormeño
Adriana Cristina Ortiz Bialous
Emilia Ottogalli
Debora Ozan
Karyna Rossana Pabes
Luis Roberto Páez
Carla Johana Pappalardo
Maia Pascual
Verónica Viviana Pavón
María Verónica Pedemonte
Mónica Peralta
Sonia Andrea Peralta
Silvina Pereyra
Natalia Pérez
Natalia Romina Pérez Castillo
Mirta Graciela Policaro
Cyntia Verónica Ponzina
Mónica Juana Quintana
Mónica Juana Quintana
Evangelina Rios
Sabrina Rivero
Silvia Rivero
María José Robles
Alejandra Roca Nieto
Cecilia Rodriguez
Roxana Karina Rodriguez
Aldana Roman
María Alejandra Romero
Cynthia Romero Vega
Natalia Alejandra Ruiz
Alicia Salinas
Mariana Sanchez
Oscar Sánchez Dardo
Pompeya De Jesús Saucedo
Roxana Savino
Roberto Sergio Sayes
Carina Andrea Schlindwein
Nora Paula Schonholz
Claudia Seery
María Rosa Selva
Virginia Ramona Sierra
María Isabel Silveira
Roberto Szurpik
Ester Taborda
Rafael Taborda
Liliana Graciela Tolay
Hilda Nancy Toledo Varas
Veronica Florencia Torre
Luis Carlos Trejo
Sara Maria Lia Trigo
Melina Evelyn Valdez
Silvia Beatriz Vazquez
María Laura Vecchioni
Ileana Vecchioni
Sandra Beatriz Vega
Amelia Vegara
Maria Gabriela Viale
Claudia Villagra
Cristina Teresa Villanueva
Sandra Analía Villarruel
Marcela Villordo
Estela Alicia Von Siebenthal
Fernanda Williner
Andrea Alejandra Zampa
Adriana Zancan
Walter Zupicich

Se terminó de imprimir en junio de 2017
en la Ciudad de Buenos Aires, República Argentina.

Aprender

Secretaría de
Evaluación Educativa

Ministerio de Educación y Deportes
Presidencia de la Nación