La educación es de todos

Mineducación

ESTABLECIMIENTO DE ESTÁNDARES DE DESEMPEÑO:

descripción de niveles y puntos de corte

icfes mejor saber

Publicación del Instituto Colombiano para la Evaluación de la Educación (Icfes)

A cargo de:
Rafael Eduardo Benjumea Hoyos
Manuel Alejandro Amado

Edición
Juan Camilo Gómez-Barrera

Diseño de portada y diagramación
Linda Nathaly Sarmiento Olaya

Fotografía original de la portada
Freepik (2019)

ISBN de la versión digital: En trámite

Bogotá, D. C., diciembre de 2020

TABLA DE CONTENIDO

LISTA DE FIGURAS, TABLAS E ILUSTRACIONES

Introducción

Los resultados de las pruebas estandarizadas involucran al menos dos aspectos básicos: el cuantitativo, relacionado con el puntaje obtenido tras responder las preguntas, y el cualitativo, concerniente a la descripción de aquello que los evaluados son capaces de hacer dadas sus respuestas. Estos dos aspectos se materializan en lo que comúnmente se ha denominado *establecimiento de estándares de desempeño (Standard Setting)*, cuyo propósito es permitir interpretar adecuadamente los resultados de una prueba con miras a un uso posterior por parte de sus usuarios: estudiantes, docentes, directivos, instituciones de educación y las entidades encargadas de la planeación y toma de decisiones en política educativa.

El término *Standard Setting* hace referencia a una metodología multifacética que comprende dos procesos: la definición de los descriptores de los niveles de desempeño, acompañado de los puntos de corte, esto da como resultado, los **niveles de desempeño**, que corresponden a categorías que describen aquello que es capaz de hacer un estudiante frente a lo que mide la prueba, de acuerdo con su puntaje.

Cizek (2012b) considera que establecer los estándares de desempeño es una de las tareas más importantes en el desarrollo, administración y proceso de reporte de resultados de pruebas estandarizadas; dicha importancia radica en las consecuencias que se derivan de las clasificaciones resultantes. En el ámbito de la educación, estas consecuencias pueden ir

desde impedir la promoción a un grado, permitir la obtención de un título, hasta ser condición para el acceso a un programa educativo u otorgar becas.

Reckase y Chen (2012) dividen la metodología de *Standard Setting* en dos partes: "(a) un medio estructurado para recolectar juicios de individuos sobre el nivel o niveles deseados de desempeño en una prueba, y (b) un proceso estadístico para convertir los juicios en un punto sobre la escala de puntajes correspondientes a cada nivel" (p. 149).

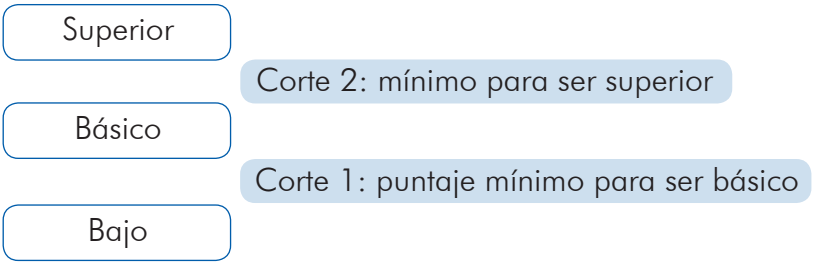
Dadas las implicaciones prácticas que pueden tener los **niveles de desempeño**, los métodos para establecerlos deben estar sustentados teóricamente, ajustarse a los propósitos de la prueba y a los argumentos que validan la decisión de generar estas categorías (Zieky & Perie, 2006). Si los niveles de desempeño se establecen con métodos sistemáticos definidos, las decisiones que se puedan tomar con base en los resultados de una prueba estarán fundamentadas.

El propósito principal de este documento es exponer una serie de pasos que permita tanto a colaboradores del Icfes como a personal externo llevar a cabo el taller de establecimiento de estándares de desempeño (*Standar Setting*) y poder generar los niveles de desempeño para las pruebas que aplica el instituto. En el primer capítulo se realizará la descripción y justificación de los dos procesos que componen el establecimiento de estándares de desempeño: primero la descripción de niveles de desempeño (DND) y luego el establecimiento de puntos de corte. En la parte final se encuentran los formatos utilizados así como una propuesta de agenda para llevar a cabo estos dos procesos.

1. Metodología para el establecimiento de estándares de desempeño

Un taller de *Standard Setting* se compone de dos procesos. El primero es la construcción de los descriptores de los niveles de desempeño (en adelante DND). Estos corresponden a afirmaciones que definen los conocimientos, habilidades y destrezas (en adelante CHD) que poseen los estudiantes en distintos niveles de apropiación (Zieky, 2012) y que son medidos en una prueba. Estas afirmaciones se clasifican en dos o más categorías (p. ej. bajo, básico, satisfactorio y avanzado), de acuerdo con su complejidad.

En segundo lugar, el establecimiento de puntos de corte, los cuales definen el puntaje mínimo requerido en una prueba para ser clasificado en cada nivel de desempeño. En términos generales, en una prueba con *n* niveles de desempeño se deberán establecer *n-1* puntos de corte. Por ejemplo, si se cuentan con tres niveles de desempeño (bajo, básico y superior), se deberán establecer dos puntos de corte en la escala de puntajes numéricos de la prueba: uno para el mínimo puntaje para ser clasificado como básico y el otro para el mínimo puntaje para ser considerado como superior.



Para establecer los estándares de desempeño existe una amplia variedad de metodologías, así como de formas de establecerlos. Una de las consideraciones principales para establecer los estándares de desempeño tiene que ver con qué se debe realizar primero, la DND o el establecimiento de puntos de corte. Aunque los dos órdenes son, en principio, admisibles, existe una robusta bibliografía que recomienda realizar la DND antes de establecer los puntos de corte en la escala numérica del puntaje (véase, por ejemplo: Perie, 2008; Egan, Schneider y Ferrara, 2012; Zieky y Perie, 2006; Plake y Reshetar, 2010; Cizek, 2012, y Cizek, G. y Buch, M., 2007). Se cuentan con al menos dos argumentos que apoyan esta recomendación. Primero, dado que la DND especifica los CHD evaluados en diferentes niveles de apropiación y dominio, es conveniente que la DND sea un insumo para el diseño de la prueba y la construcción de sus ítems (preguntas) (Perie, 2008). Segundo, si se realiza la DND posterior al establecimiento de puntos de corte, existe el riesgo de que los CHD definidos en los niveles no se correspondan con aquellos que se pretende evaluar, lo cual pone en riesgo la validez de la prueba (Egan *et al*, 2012: 79).

Por estas razones, la metodología adoptada por el Icfes para el establecimiento de estándares de desempeño procederá, primero, con la DND y, luego, con la determinación de puntos de corte.

1.1 Método para la descripción de niveles de desempeño (DND)

En el caso de las evaluaciones educativas, es conveniente integrar la DND con el diseño de las pruebas. Si la DND se realiza en una etapa temprana o paralela al diseño de una prueba, esta deberá convertirse en un insumo para la construcción de las preguntas. Esto se debe a que, al tener previamente la cantidad y la exigencia de cada nivel, las preguntas podrán elaborarse de tal forma que permitan diferenciar los niveles de desempeño. De acuerdo con esto, se describe el modelo de diseño de pruebas utilizado en el Icfes y cómo este se integra con la DND.

La DND involucra tres pasos: primero, especificar el número y las etiquetas de los niveles; segundo, la definición de las descripciones generales de los niveles (nombradas por algunos autores como *policy definitions*), y, tercero, complementar la descripción general con descriptores específicos para cada nivel y para cada prueba de la evaluación. Cabe indicar que los pasos de la DND requieren la participación de los encargados de las políticas educativas (o contratantes de una evaluación). Se trata de expertos disciplinares elegidos porque, en definitiva, la DND refleja las políticas educativas enmarcadas en etiquetas como "satisfactorio", así como las expectativas de los CHD evaluados en la prueba (Perie, 2008). A continuación, se describen cada uno de los tres pasos.

1.1.1 Número de niveles y etiquetas

Los encargados de las políticas educativas (o contratantes de una evaluación) son los actores principales en la selección de la cantidad de niveles requeridos. Idealmente, se debe elegir la menor cantidad de niveles de desempeño que permitan cumplir los propósitos de un examen. Por ejemplo, en una prueba de selección de candidatos, lo recomendable es tener solo dos niveles de desempeño: aprobado y reprobado. De acuerdo con Perie (2008), una cantidad mayor de niveles de desempeño dificulta establecer diferencias significativas entre los niveles. Adicionalmente, mientras mayor sea el número de niveles, más grande será el trabajo requerido para la DND y el establecimiento de puntos de corte. En exámenes de Estado, por ejemplo, esto podría dificultar su manejo en términos de tiempo y recursos requeridos.

De acuerdo con los usos que se dan a las evaluaciones educativas a nivel mundial, no se suelen plantear más de cuatro niveles de desempeño. Leyes como "Que ningún niño se quede atrás" (NCLB, por sus siglas en inglés), en Estados Unidos, exigen que las evaluaciones que realicen los Estados reporten al menos tres niveles de desempeño, uno para satisfactorio, un nivel por debajo y otro por encima. Sin embargo, la mayoría de Estados reportan cuatro niveles de desempeño, lo que permite diferenciar mejor aquellos estudiantes que están cerca de ser satisfactorios y aquellos que están lejos de serlo.

En el caso de algunos exámenes realizados por el Icfes, se han reportado tres o cuatro niveles de desempeño, en consonancia con el decreto 1290 de 2009 del Ministerio de Educación Nacional (en adelante MEN). Allí se pide a los establecimientos de educación básica y media generar una escala de valoración de los desempeños de los estudiantes con cuatro niveles, con el fin de facilitar la movilidad de los estudiantes entre establecimientos educativos.

Una vez determinada la cantidad de niveles de desempeño, el paso siguiente es darles un nombre o etiqueta. Por ejemplo, algunas etiquetas pueden ser aprobado, reprobado, bajo, básico, satisfactorio, avanzado, superior etc. Cizek y Bunch (2007) recomiendan que estas etiquetas sean elegidas cuidadosamente, ya que se deben relacionar con el propósito de la evaluación, el dominio evaluado y las inferencias que se pretenden establecer a partir de los resultados. Zieky, Perie y Livingston (2008: 21) recomiendan que, de ser posible, se debe considerar "el uso de etiquetas neutrales como *Nivel 1, Nivel 2, y Nivel 3*, para los niveles de desempeño. Las etiquetas neutrales evitan el exceso de significado que con frecuencia se les da a etiquetas más descriptivas". Esta última recomendación ha sido acogida por el Icfes, como veremos en el ejemplo presentado en el apartado 1.2.2.

1.1.2 Definición de descripciones generales de los niveles de desempeño

Las descripciones generales de los niveles de desempeño, conocidas en algunos contextos como *policy definitions* (Perie, 2008: 16), son definiciones genéricas que aplican para todas las pruebas que componen una evaluación, independientemente de los dominios que evalúen. Son afirmaciones generales que reflejan la posición de los encargados de la política educativa en cuanto al grado deseado de desempeño en cada nivel. Por ejemplo, básico puede significar la superación de los desempeños necesarios en relación con las áreas obligatorias y fundamentales, teniendo como referente los estándares básicos, las orientaciones y lineamientos expedidos por el MEN (decreto 1290, 2009). Estas definiciones deberían ser consistentes a través de todos los grados y todas las áreas, lo que resulta muy útil para asegurar el mismo nivel de exigencia para cada nivel de desempeño por grado y área.

De acuerdo con Perie (2008), para un determinado examen, una definición general debe realizarse para cada nivel de desempeño sin importar el grado o el área de contenido. Esto no aplica para el nivel más bajo, pues este debe reflejar un rendimiento por debajo del mínimo requerido para alcanzar el nivel inmediatamente posterior. Así, en una prueba con dos niveles de desempeño, solo es necesario definir el nivel para los que aprueban, ya que los que no aprueban quedan definidos por defecto.

A manera de ejemplo, se presentan las definiciones generales obtenidas para el examen Saber 3.°, 5.° y 9.° en un taller realizado en diciembre de 2019 que incluía las pruebas de Matemáticas, Ciencias Naturales y Educación Ambiental, Competencias Comunicativas: Lenguaje y Pensamiento Ciudadano.

Definición general de niveles

Nivel 3: El estudiante muestra un desempeño satisfactorio de aquellos CHD descritos en los estándares (o marco teórico de la prueba) que son evaluados.

Nivel 2: El estudiante muestra un desempeño que indica un nivel aceptable de desarrollo de los CHD descritos en los estándares (o marco teórico de la prueba).

Nivel 1: El estudiante muestra un desempeño básico que indica un desarrollo limitado de algunos de los CHD descritos en los estándares (o marco teórico de la prueba).

Por debajo de 1: El estudiante podría manifestar algunas habilidades simples, pero no hay evidencia para afirmar que alcance un desarrollo básico de alguno de los CHD descritos en los estándares (o marco teórico de la prueba).

1.1.3 Descripciones específicas de niveles

Las descripciones específicas son afirmaciones que expresan los CHD requeridos para alcanzar cada nivel de desempeño y están directamente ligadas a una prueba. Estos descriptores deben proveer información sobre lo que un estudiante sabe y puede hacer. Además, pueden indicar lo que le falta para alcanzar el nivel siguiente. La construcción de las descripciones específicas se puede efectuar a partir del dominio de la prueba o de las especificaciones de prueba. Las preguntas pueden ayudar a esta descripción, dependiendo del momento en que se realice el taller de la DND. Sin embargo, se debe evitar describir los CHD específicos para responder una pregunta particular. Por el contrario, se deben describir los CHD para responder un conjunto de preguntas similares.

En el caso de las pruebas que aplica el Icfes, las descripciones específicas de los niveles se realizarán a partir de las especificaciones de prueba. En aras de la claridad del método, a continuación, se expone cómo el Icfes diseña la mayoría de las pruebas que aplica.

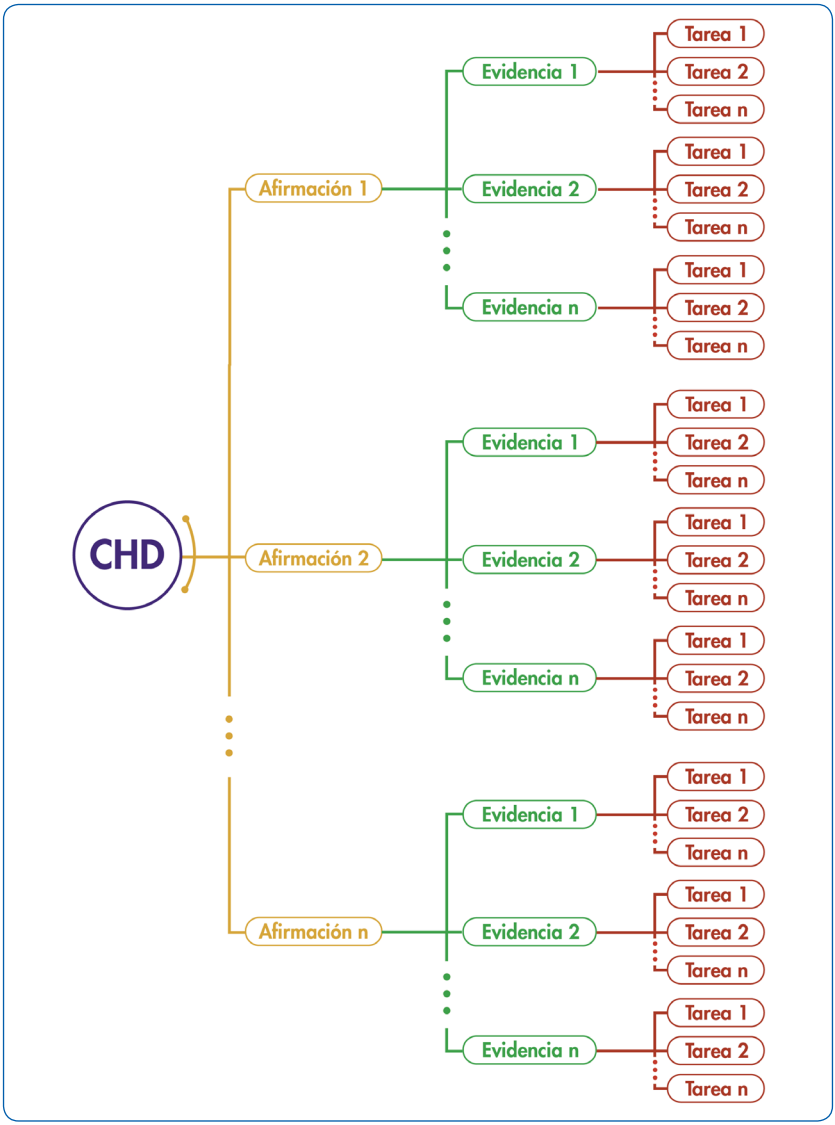
Las pruebas que aplica el Icfes se fundamentan en el modelo de diseño centrado en evidencias (en adelante DCE). Este método tiene como objetivo garantizar que exista una línea clara de razonamiento entre las respuestas que dan los estudiantes a las preguntas y las afirmaciones que, a partir de estas respuestas, se pretenden hacer sobre los CHD evaluados en la prueba (Mislevy, 2003, 2017). Dado que estas afirmaciones deben estar soportadas por evidencias suficientes e independientes,

el DCE propone que, para recoger estas evidencias, cada pregunta de la prueba demande el cumplimiento de una determinada tarea: una situación específica en la que un estudiante pueda exhibir parte de los CHD evaluados.

De acuerdo con la anterior, el DCE propone que una prueba se estructure en términos de tres aspectos: 1) las *afirmaciones* sobre los CHD que constituyen el dominio de la prueba, es decir, lo que se pretende medir; 2) las *evidencias* o aspectos de la conducta observable de los estudiantes que soportan las afirmaciones, y 3) las *tareas*, que describen las acciones y situaciones concretas que permiten a los estudiantes evidenciar los CHD que evalúa la prueba. A partir de las tareas se construyen las preguntas de una prueba (Icfes, 2018).

Los tres aspectos descritos se relacionan de manera jerárquica: cada tarea indaga por una habilidad, conocimiento o destreza específicos de tal forma que diferentes tareas indagarán por diferentes CHD. Cada grupo de tareas permitirá recoger una determinada evidencia que, a su vez, soportará una determinada afirmación sobre los CHD que constituyen el dominio de la prueba. Esta estructura jerárquica de afirmaciones, evidencias y tareas constituye las así llamadas especificaciones de prueba. La figura 1 representa la estructura jerárquica que asumen las especificaciones:

Figura 1. *Estructura de las especificaciones de una prueba de acuerdo con el DCE*



Dado que los CHD medidos están consignados en las especificaciones de las pruebas, es razonable que la definición de niveles de desempeño sea extraída de las especificaciones, puntualmente, de las afirmaciones y evidencias de la prueba. Los pasos del método para realizar las descripciones específicas de niveles por prueba, a partir del DCE, están fundamentados en los trabajos de Perie (2008) y Plake *et al*. (2010) y se exponen a continuación:

1. Se conforma un comité de jueces que tendrán como objetivo establecer las descripciones de los niveles para cada prueba particular. Aunque en la literatura no se especifica un número determinado de jueces, es deseable lograr representación en términos de género, condición socioeconómica de la población que atienden, tipo de institución (pública, privada) y grupos minoritarios. Perie (2008: 19) propone entre cinco y no mucho más de ocho jueces. En pruebas piloto de DND, el Icfes ha encontrado que diez jueces pueden satisfacer la demanda de representación. El perfil mínimo de los jueces incluye:

- Expertos curriculares en el área evaluada.
- Profesores en ejercicio que conozcan la población.
- Otros responsables de las políticas educativas.

2. Se hace una presentación del DCE y de las especificaciones de prueba hasta el nivel de evidencias.

3. A todos los integrantes del comité se les orienta sobre la definición y objetivos del establecimiento de estándares,

específicamente, de la DND. En esta presentación es importante incluir ejemplos de niveles de desempeño de otras pruebas (aplicadas a poblaciones cercanas a la población objetivo) y la distribución de la población, con el propósito de ejemplificar el producto esperado del taller, sensibilizar a los jueces con respecto a su percepción del comportamiento de la población y sus expectativas frente al desempeño de esta (lo que contaría como un desempeño característico de cada uno de los cuatro niveles).

4. Se desarrolla el taller de la descripción de los niveles de desempeño específicos en dos momentos y con los siguientes insumos (los primeros tres insumos deben enviarse a los jueces previo al taller):

- Marco de referencia o guía de orientación de la prueba.
- Cuadernillo de práctica (ítems liberados de la prueba objeto del taller, si los hay).
- Presentaciones del taller (diapositivas en las que se presentan los temas que se desarrollarán en el taller).
- Especificaciones: afirmaciones y evidencias[1].
- Sistema de clasificación de objetivos cognitivos: taxonomía de Bloom (un cuadro que sintetiza este sistema se anexa a este documento).

1 No necesariamente se exponen las tareas, pero si alguna evidencia resulta abstracta o la clase de tareas a las que se asocia no resulta clara, el facilitador presentará las tareas durante el taller (no antes) o presentará de manera general los contenidos conceptuales asociados a las tareas.

Primer momento de la descripción de niveles de desempeño específicos: ordenamiento de evidencias de la prueba.

Reunidos en dos grupos de alrededor de cinco personas, se lleva a cabo un **ordenamiento de las evidencias de la prueba**. Este orden se hará a manera de un continuo de evidencias de menor a mayor complejidad cognitiva. Dado el caso en que una misma evidencia agrupe tareas de complejidad muy diferente, se generarán dos o más descripciones de la evidencia (descriptores) que den cuenta de la diversidad en su dificultad; estos descriptores ocuparán diferentes posiciones en el ordenamiento.

Los descriptores provenientes de una misma evidencia pueden obtenerse, al menos, de dos maneras: primero, las descripciones pueden variar en cuanto a **aspectos del contenido conceptual de las tareas** asociadas a una evidencia. Por ejemplo, considere la siguiente evidencia de las especificaciones de una prueba de comprensión de lectura:

> **E.1.** Identifica elementos del contenido de diferentes textos narrativos (tiempo, lugares, eventos, personajes y narrador).

Esta evidencia puede estar conformada por tareas de distinta complejidad, por lo que conviene generar al menos dos descriptores diferentes que se ubiquen en distintos puntos del ordenamiento. Estas dos descripciones pueden ser:

E.1.a. Identifica elementos del contenido de diferentes textos narrativos que responden a preguntas del tipo: ¿cómo?, ¿cuándo?, ¿dónde?, ¿cómo? y ¿por qué?

E.1.b. Identifica elementos del contenido de diferentes textos narrativos a partir de descripciones presentes (identifica personajes principales, personajes secundarios, protagonistas, antagonistas, rasgos físicos o psicológicos).

De este modo, las nuevas descripciones, **E.1.a** y **E.1.b**, ocuparán posiciones diferentes en el ordenamiento de evidencias para, así, dar cuenta de la diversa complejidad de las tareas asociadas a la evidencia **E.1** (que ya no aparecerá en el ordenamiento en su forma original).

La segunda forma en la que se pueden obtener diferentes descripciones de una misma evidencia consiste en modificar los verbos de objeto cognitivo que aparezcan en la redacción original de la evidencia. Para que la modificación respete el ordenamiento en términos de la complejidad, se puede usar un sistema de clasificación de objetivos cognitivos como la taxonomía de Bloom. Concretamente, se pueden emplear los verbos indicadores de cada una de las categorías de habilidades de pensamiento que se definen en esta taxonomía (como se menciona arriba, la taxonomía de Bloom es parte de los insumos que se entregan a los jueces antes del desarrollo del taller con el propósito de familiarizar dicha clasificación). Un ejemplo de esta estrategia de modificación, adaptado de Plake *et al*. (2010), es el siguiente:

> **A.1.** El estudiante **evalúa** los modos en que los desarrollos y procesos históricos concretos se relacionan con procesos regionales, nacionales o globales más amplios.

La evidencia original emplea el verbo de habilidad cognitiva *evaluar*. Dicho verbo se encuentra, en términos de la taxonomía de Bloom, en un grado de complejidad superior. Dado que varias habilidades cognitivas de diferente complejidad pueden ser asociadas al objeto del verbo ("los modos en que los desarrollos y procesos históricos concretos se relacionan con procesos regionales, nacionales o globales más amplios"), se pueden generar uno o más descriptores diferentes que den cuenta de estas diferentes habilidades cognitivas:

> **A.1.a.** El estudiante **explica** los modos en que los desarrollos y procesos históricos concretos se relacionan con procesos regionales, nacionales o globales más amplios.

> **A.1.b.** El estudiante **nombra** los modos en que los desarrollos y procesos históricos concretos se relacionan con procesos regionales, nacionales o globales más amplios.

En este caso, de acuerdo con la taxonomía de Bloom, el descriptor de la evidencia original **A.1** puede mantenerse en el ordenamiento en una posición superior al nuevo descriptor **A.1.a** que, a su vez, ocupará una posición superior al nuevo descriptor **A.1.b**. Para completar el ordenamiento se sugiere el siguiente procedimiento:

a. En el primer momento los jueces se dividen en dos grupos para plantear de forma separada una versión del ordenamiento. El número de descriptores obtenidos no debe ser menor que el número de evidencias de la prueba ni corresponder exactamente a las tareas (no necesariamente los ítems que aparecen en una forma representan a todas las tareas de las especificaciones de la prueba). Además, se recomienda que el número de descriptores que se deriven de una sola evidencia no sea mayor que el número de niveles de desempeño. Así, por ejemplo, si una prueba para la que se han definido cuatro niveles de desempeño tiene diez evidencias y 40 tareas, el número de descriptores obtenidos no puede ser menor que diez; además, por cada una de las diez evidencias, se recomienda no generar más de tres descriptores (correspondientes a los cuatro niveles de la prueba menos el primer nivel).

b. Después de que cada grupo propone un ordenamiento inicial, ambos grupos se reúnen y compararan las descripciones y los ordenamientos. Para ello, se debe consignar en una tabla el número de descripciones resultantes y el código de las evidencias que los integran, según las especificaciones de la prueba (ver tabla 1).

c. A partir de la comparación, se obtiene un ordenamiento final que sintetiza los aportes de ambos grupos. Para obtenerlo, se señalan las coincidencias, si las hay, en la posición que los dos grupos le hayan asignado a un descriptor. Dado el caso en que existan discrepancias con

respecto a la posición de un descriptor determinado, se indaga por las razones que cada grupo tiene para ubicar a ese descriptor en una u otra posición, y se evalúan estas razones para llegar a un consenso.

d. Finalmente, una vez se sintetice el listado de los descriptores de evidencias, se enumeran de 1 en adelante (siendo 1 la descripción de la evidencia menos compleja del ordenamiento) y se redactan en limpio los descriptores resultantes, eliminando los códigos de las evidencias de los que se obtienen. El ordenamiento preliminar será un insumo del segundo momento en el proceso de la DND.

A manera de ejemplo, se pueden ver los ordenamientos de una prueba hipotética propuestos por dos grupos de jueces. Los descriptores (D1 a D10) están ordenados en términos de complejidad, según el criterio de los jueces, en donde el descriptor en posición 1 es el menos complejo y el descriptor en la posición 10, el más complejo. Los descriptores sombreados corresponden a las coincidencias en la posición de los descriptores. Dichos descriptores se anotan en la respectiva columna de coincidencias:

Tabla 1. *Ordenamientos propuestos por los dos grupos de jueces*

Ordenamiento Grupo A	Coincidencias	Ordenamiento Grupo B
1. D1	D1	1. D1
2. D3		2. D10
3. D5	D5	3. D5
4. D7	D7	4. D7
5. D10		5. D3
6. D2	D2	6. D2
7. D4		7. D6
8. D9		8. D9
9. D6	D4	9. D4
10. D8	D8	10.D8

Para resolver los desacuerdos en la posición de un descriptor, se indaga si el descriptor en cuestión es más (o menos) complejo que alguno de los descriptores sobre los que ya existe un acuerdo. Por ejemplo, de acuerdo con la tabla 1, no hay acuerdo sobre la ubicación concreta de D4, pues el grupo A lo ubica en la posición 7 mientras que el grupo B lo ubica en la posición 9. No obstante, para ambos grupos D4 es más complejo que D2 y menos complejo que D8. Como D4 está entre D2 y D8, se ubica a D4 entre dichos descriptores en la columna de coincidencias sin asignarle un número de posición concreto como se subraya en la tabla 1. Este procedimiento

se repite hasta que todos los descriptores propuestos reciban una posición en el ordenamiento que el grupo total de jueces apruebe.

En el caso de que la cantidad de descriptores no coincida en los dos grupos, es decir, que un grupo haya creado más divisiones de evidencias que el otro o que los descriptores no coincidan, deben discutirse qué elementos de las evidencias llevaron a esas discrepancias y poder llegar a acuerdos y continuar como se expresó en el párrafo anterior. Es importante anotar que, aunque existan divisiones de evidencias, estas pueden fusionarse de nuevo en el **segundo momento**, porque será posible asignarlas al mismo nivel de desempeño, de acuerdo con los pasos que veremos a continuación.

Segundo momento de la descripción de niveles de desempeño específicos: asignación de grupos de descripciones de evidencia a los niveles de desempeño.

A partir del ordenamiento de las descripciones de las evidencias en términos de complejidad en el primer paso, se asignan grupos de estas descripciones a cada uno de los cuatro niveles de desempeño. La asignación resultante debe ser **jerárquica** e **inclusiva**: entre mayor sea el nivel, mayor será la complejidad de las habilidades exhibidas y cada nivel será, en parte, la conjunción de los niveles anteriores.

Para asignar grupos de evidencias a niveles, los jueces se organizarán en dos grupos diferentes a los grupos del primer momento. Los pasos básicos en este segundo momento se describen a continuación:

a. Se identifican los descriptores que, a juicio del grupo, describen los CDH que ubican a un estudiante en el nivel de desempeño correspondiente a aquel que cumple con las expectativas de la política educativa. Por lo general, este nivel es etiquetado como satisfactorio. En el caso del ejemplo presentado en la tabla 2 abajo, se trata del nivel 2. Luego, se realiza el mismo ejercicio con el nivel de desempeño anterior, de manera que la descripción de estos niveles sea congruente con las descripciones del nivel 2. Esto significa que cada nivel *N* incluirá los descriptores del nivel *N-1* más otros descriptores de posiciones superiores en el ordenamiento.

Por ejemplo, si al nivel 1 se asociaron los descriptores que ocupan las posiciones 1 a 5 del ordenamiento de evidencias, al nivel 2 se le asignarán esos mismos descriptores más otros descriptores del continuo que ocupen posiciones mayores o iguales a 6. Similarmente, el nivel 3 tendrá a todas aquellas evidencias que componen los niveles 1 y 2, más otras evidencias en posiciones superiores en el ordenamiento. Al final, el último nivel tendrá asignado todo el continuo de los descriptores definido en el primer momento. De esta forma, se recoge la idea de que los niveles de desempeño son **jerárquicos** e **inclusivos**.

b. Los diez jueces se reúnen de nuevo para sintetizar sus resultados. La asignación final de descripciones a cada uno de los niveles terminará la tarea de la DND. Si existe desacuerdo con respecto a las descripciones

que definen un nivel particular, se discuten las razones del desacuerdo. Si el desacuerdo persiste, el grupo de evidencias que defina el nivel será la media de las dos posiciones que cada grupo de jueces le asigne al último descriptor que ocupa el nivel. Si el resultado de la media no es un número entero, se acercará al menor número entero más cercano a la media. Así, por ejemplo, si para un grupo de jueces el orden de descriptores asociado al nivel 2 termina en la posición seis del ordenamiento y, para el otro grupo, termina en la posición nueve, el nivel 2 terminará en el descriptor con la posición siete.

c. En la última actividad del taller de DND, los panelistas completan una evaluación donde califican qué tanto entendieron el trabajo que debían realizar, la efectividad de la metodología para alcanzar los objetivos del taller y sus recomendaciones de mejora. Este ejercicio permite la identificación de vacíos que puedan existir en la explicación del método, así como la calidad de la instrucción recibida por parte de los profesionales del área y los moderadores de los grupos de jueces.

Cabe resaltar que las descripciones específicas de niveles que resulten de este taller están sujetas a cambios durante las discusiones que se adelanten en la ronda 3 del taller de establecimiento de puntos de corte, una vez se contrasta el ordenamiento de descriptores con un cuadernillo de preguntas ordenados. La tabla 2 ilustra una posible descripción de niveles para una prueba de lenguaje con cuatro niveles de desempeño:

Tabla 2. *DND para la prueba de competencias comunicativas: Lenguaje 5.°*

Nivel	Descriptores
Por debajo de 1 El estudiante en este nivel:	Podría identificar el significado de algunas palabras que aparecen en un texto o algunos elementos de textos narrativos como tiempos, lugares, hechos y personajes.
Nivel 1 El estudiante en este nivel:	Identifica elementos del contenido de diferentes tipos de textos (tiempo, lugares, hechos, personajes y narrador).
	Identifica el significado de palabras que aparecen de manera explícita en el texto.
Nivel 2 Además de lo descrito en el nivel anterior, el estudiante en este nivel:	Describe elementos no lingüísticos de textos discontinuos narrativos (por ejemplo, convenciones universales en caricaturas o historietas).
	Identificar actos de habla directos (reconoce cuando un autor o persona hace una afirmación, una pregunta, una petición, etc.)

Continúa

Nivel	Descriptores
	Reconoce el tema o el problema abordado en un texto.
	Identifica la función de marcadores y conectores que aparecen de manera explícita en el texto (por ejemplo, conectores como "sin embargo", "además", "porque", "entonces" y marcadores como "en primer lugar", "había una vez" y "en consecuencia").
	Identifica las funciones de las partes en las que se estructura un texto.
	Reconoce resúmenes y paráfrasis apropiados de un texto.
	Identifica la relación entre las personas que desempeñan un papel en una narración (por ejemplo, relaciones entre personajes, narrador y autor).

Continúa

Nivel	Descriptores
Nivel 3 Además de lo descrito en el nivel anterior, el estudiante en este nivel:	Identifica la relación entre las personas que desempeñan un papel en una secuencia argumentativa o dialógica (por ejemplo, identifica cuando un personaje o autor contradice, apoya o es una fuente de otro personaje o autor).
	Establece relaciones entre elementos lingüísticos y no lingüísticos en textos discontinuos explicativos y argumentativos (por ejemplo, textos que contienen diagramas, esquemas, líneas de tiempo, entre otros).
	Reconoce tesis, razones a favor o razones en contra presentes en un texto argumentativo.
	Identifica estrategias discursivas del texto; por ejemplo, identificar el propósito comunicativo del autor del texto y si emplea estrategias argumentativas como ejemplificación y cita de autoridades.

Continúa

Nivel	Descriptores
	Establece relaciones entre el texto y el contexto. Por ejemplo, reconoce la audiencia a la que se dirige el texto y en qué tipo de publicación se encontraría.
	Establece relaciones entre diferentes textos; por ejemplo, reconoce cuando dos textos tratan del mismo tema, pero tienen diferentes posturas.

1.2 Método para establecimiento de puntos de corte

Los métodos para establecer puntos de corte son diversos (véase, por ejemplo, Lane *et al*, 2016; Aranguren y Hoszowski, 2017; Cizek, 2012; Herrera *et al*, 2009; Kingston y Tiemann, 2012; Perie y Thurlow, 2012, y Peterson *et al*, 2011). No obstante, estos métodos tienen en común la necesidad de contar con un grupo de jueces, quienes establecerán los puntos específicos que, finalmente, determinarán las escalas en el puntaje numérico de la prueba correspondientes con la descripción cualitativa obtenida en la DND. De acuerdo con la clasificación adoptada por Aranguren y Hoszowski (2017), los métodos para establecer puntos de corte se pueden organizar en cuatro grupos, en términos del fundamento que emplean los jueces para asignarlos:

- **Tipo A.** Juicios basados en la revisión de los ítems (y sus indicadores estadísticos).
- **Tipo B.** Juicios a partir del trabajo realizado por los examinados.
- **Tipo C.** Juicios basados en los perfiles de rendimiento.
- **Tipo D.** Juicios basados en los evaluados o candidatos.

A continuación, para contextualizar el método que utiliza el Icfes al establecer los puntos de cortes, y apoyar la elección de *Bookmark* frente a los otros, se revisan los métodos que han tenido mayor uso dentro de las cuatro categorías mencionadas anteriormente. Dados los fundamentos teóricos que sustentan el diseño, construcción, aplicación y calificación de estas pruebas, se mostrará que, de acuerdo con los criterios que

sugiere la literatura, el método *Bookmark* resulta ser uno de los más convenientes para establecer los puntos de corte de los instrumentos desarrollados por el Icfes.

Métodos Tipo A

En estos métodos no se examinan directamente las respuestas de los estudiantes, sino el contenido y, de ser posible, los datos psicométricos de los ítems. Entre este grupo, los métodos *Angoff* y *Bookmark* son los más empleados. En ambos métodos se exige, aunque de manera diferente, que los jueces emitan veredictos sobre la probabilidad que tiene un **estudiante frontera** o **mínimamente competente** de responder correctamente una pregunta, es decir, un estudiante del que se espera que tenga un **dominio mínimo** de todos los CHD que lo ubican en un nivel de desempeño *N*, pero que no ha logrado aún todos los CHD que lo ubicarían en el nivel inmediatamente superior (*N*+1).

Método Angoff y sus modificaciones.

En este método se reúne un panel de jueces a los que se les solicita que revisen el contenido de un número determinado de ítems, para que estimen la probabilidad de que un *estudiante frontera* responda correctamente a cada uno de estos ítems. Lo anterior convierte al método Angoff en un método costoso en términos de recursos cognitivos y tiempo, pues exige a los jueces emitir estimaciones de probabilidad que en muchas ocasiones no son precisas o no están capacitados para hacer. Adicionalmente, el método es de difícil aplicación para pruebas que incluyan preguntas abiertas o politómicas

(Aranguren y Hoszowski, 2017). Sin embargo, se han hecho mejoras a este método con el objetivo de fortalecer sus puntos débiles. Entre estas modificaciones se encuentran:

- Incluir procesos iterativos de revisión y discusión para llegar a consensos entre los jueces, de manera que se aumente el grado de acuerdo.
- Entregar información estadística de los ítems (p. ej. proporción de respuestas correctas para cada ítem) para promover el acuerdo entre los jueces.
- Cambiar la instrucción para que, en lugar de estimar la probabilidad de responder correctamente el ítem, los jueces solo indiquen, en términos de "sí" o "no", si un estudiante frontera respondería correctamente el ítem, lo que permite simplificar la tarea de los jueces.

Método Bookmark.

En este método, un grupo de jueces revisa un conjunto de ítems que componen la prueba, los cuales están organizados dentro de un cuadernillo de acuerdo con la habilidad requerida para contestarlo correctamente de menor a mayor. Dicha habilidad es determinada mediante la aplicación de la teoría de respuesta al ítem (TRI). El uso del cuadernillo permite a los expertos tener una visión global del grado de complejidad de la prueba y ayuda a que centren su atención en aquellos ítems particularmente sensibles para los estudiantes frontera (Herrera Ortiz *et al*, 2009). En términos generales, la tarea de los expertos consiste en fijar una marca en el cuadernillo para cada punto de corte. Esto significa que, para cuatro niveles

de desempeño, se pondrán tres marcas en el cuadernillo. La posición de las marcas dependerá de lo que, según los jueces, puedan hacer los estudiantes frontera (mínimamente competentes). De este modo, los jueces tendrán que imaginar tantos estudiantes mínimamente competentes como marcas se requieran.

La tarea de los jueces en el método *Bookmark* consiste en determinar cuáles son las preguntas que cada uno de los estudiantes mínimamente competentes, para ser ubicados en un nivel, podría contestar correctamente con una probabilidad constante conocida como *probabilidad de respuesta* (PR). La probabilidad de respuesta más empleada es 2/3 (o 0,67); pero otros valores, desde 0,5 a 0,8, también pueden emplearse. La elección de la probabilidad de la respuesta depende principalmente de los conceptos de información total del ítem e información de la respuesta correcta del ítem. La información total del ítem se maximiza cuando la PR se acerca a 0,5, mientras que la información de la respuesta correcta se maximiza cuando PR se acerca a 0,67 (2/3). Para argumentos acerca de cuál de estos dos conceptos de información debe preferir el diseñador de pruebas en la elección de una PR, se puede consultar Huynh (2006).

De acuerdo con Cizek (2012), se recomienda usar un valor de PR cercano a 0,5 para pruebas cuya dificultad sea alta, pues mayores valores de PR impondrían un reto mayor en estas pruebas para los estudiantes de cualquier nivel, dado que valores altos de PR suelen dar como resultado marcas

altas[2]. De este modo, por ejemplo, si los jueces deciden que un estudiante mínimamente competente que está en el nivel 2 puede responder correctamente con una PR de 2/3 hasta el ítem en la posición 10 del ordenamiento, una marca será puesta en este ítem (Herrera Ortiz *et al*., 2009, p. 45). El punto de corte se obtendrá a partir de la habilidad (estimada en el modelo TRI) asociada a dicho ítem.

En síntesis, *Bookmark* es uno de los métodos más utilizados en la actualidad para la determinación de los puntos de corte en las evaluaciones del ámbito educativo. Se ha evidenciado ampliamente su uso en la mayoría de los estados norteamericanos, y se han reportado resultados satisfactorios en Perú, Chile, Guatemala, Brasil y México (Aranguren y Hoszowski, 2017). Además, el método puede usarse en pruebas que contengan tanto ítems cerrados de opción múltiple con única respuesta como ítems abiertos o de respuesta construida.

Métodos Tipo B

Los métodos de esta categoría son *holísticos*. Esto significa que los jueces revisan y establecen juicios a partir de las respuestas completas a varios ítems de una muestra de los evaluados.

2 En este documento se empleará el valor de 2/3 para la probabilidad de respuesta (PR). Sin embargo, cabe aclarar que en caso en el que se privilegie la información de la respuesta correcta, y el modelo de TRI aplicado sea 3PL, se recomienda emplear un valor de PR igual a (2 + c)/3, donde c es el parámetro de pseudo-azar Huynh (2000).

Esto es posible porque se asume que las habilidades, aunque conceptualmente distinguibles, en la práctica, están ampliamente integradas en el desempeño total del estudiante (Cizek, 2012: 8).

Método BoW.

El método BoW (*Body of Work method*, en inglés) es un método holístico utilizado normalmente con pruebas de preguntas de respuesta construida en las que se pide a los estudiantes redactar textos de estructura compleja como textos argumentativos, explicativos y narrativos. La aplicación de este método sigue cinco pasos: (1) definición de los niveles de desempeño; (2) selección de los trabajos de los estudiantes de una muestra representativa por evaluar y categorizar; (3) selección y capacitación de los jueces que participarán del establecimiento de puntos de corte; (4) correspondencia entre el desempeño de los estudiantes y los niveles de desempeño establecidos, y (5) establecimiento de puntos de corte (Aranguren y Hoszowski, 2017).

Las preguntas de la prueba y las respuestas dadas por los estudiantes deben presentarse en un cuadernillo llamado "cuadernillo de respuestas". La tarea de los jueces consiste en revisar este cuadernillo y analizar el desempeño de cada estudiante, indicando en cada cuadernillo el nivel de desempeño alcanzado. Así mismo, los jueces deben justificar el nivel alcanzado, estableciendo, para cada caso, cuáles son los CHD observados en las respuestas.

Método del juicio analítico.

En el método del juicio analítico de Plake y Hambleton (1998, 2001), así como en el método BoW, se pide a los jueces que clasifiquen el trabajo realizado por los estudiantes, teniendo en consideración los distintos niveles de desempeño definidos. La diferencia con el método anterior es que, en este, la tarea de los jueces se descompone en partes, y en vez de evaluar el material de los estudiantes en su conjunto, los jueces deben evaluar el desempeño de los estudiantes, revisando una tarea a la vez (Aranguren y Hoszowski, 2017). Al descomponer el trabajo en distintas etapas, el proceso suele requerir menos tiempo que cuando se evalúa todo el trabajo en una única instancia.

Métodos Tipo C

Estos métodos dependen de juicios basados en los perfiles de rendimiento y, por tanto, son holísticos. La diferencia con respecto a los métodos tipo B es que la noción central es la de *perfil de desempeño*, entendida como una caracterización del desempeño mínimamente aceptable en una evaluación.

Método del perfil dominante.

Es un método basado en los perfiles de desempeño y resultados típicos. Acá, primero, se divide la prueba en componentes que midan diferentes conocimientos y habilidades; posteriormente, se definen perfiles de desempeño para cada uno de los componentes de la prueba. A partir de dichos perfiles, se

determina un punto de corte para cada uno de los componentes de forma separada. Ese punto de corte se obtiene luego de especificar los criterios que se consideran más adecuados para diferenciar los puntajes que cumplen un perfil aceptable de aquellos que no lo cumplen (Kingston y Tiemann, 2012, p. 222).

Método de la captura de políticas.

Similar al método del perfil dominante, la prueba es dividida en partes o componentes que evalúan diferentes habilidades, para luego obtener los perfiles de desempeño de cada componente. En el método de la captura de políticas, los jueces asignan individualmente una calificación numérica al perfil de desempeño global mínimamente aceptable. Más adelante, en una fase de análisis de datos, se definen los pesos relativos que los jueces asignan implícitamente a cada parte de la prueba al asignar la calificación numérica global.

A diferencia del método del perfil dominante, en este método no se busca que los jueces hagan explícitos los criterios que emplean para diferenciar los puntajes que cumplen un perfil aceptable de aquellos que no lo cumplen. La calificación dada por los jueces se establece individualmente en una primera ronda, luego se discute con el grupo total de jueces y se ajustan individualmente en una segunda ronda para ser confirmadas o modificadas (Kingston y Tiemann, 2012).

Métodos Tipo D

Estos métodos dependen de los juicios sobre las respuestas y habilidades que exhibe un grupo de evaluados. En esa medida, estos requerimientos implican que estos métodos demanden, para su aplicación, una cantidad amplia de tiempo (Perie y Thurlow, 2012). Uno de los métodos más representativos de este tipo es el de contraste de grupos:

Método de contraste de grupos.

Para aplicar este método, es necesario tener acceso a una muestra de resultados de los evaluados, los cuales se dividen en dos grupos, tomando como insumo los juicios sobre sus conocimientos y habilidades. Un grupo estará conformado por evaluados *satisfactorios* (los que pasan), y el grupo de contraste estará conformado por evaluados *no satisfactorios* (los que fallan). En cada nivel de puntuación, los jueces deben estimar la probabilidad de que un evaluado sea miembro del grupo satisfactorio. Si los jueces creen que es igual de perjudicial pasar a un miembro del grupo que no es satisfactorio (esto se conoce como *falso positivo*) que hacer que un miembro del grupo satisfactorio falle (*falso negativo*), el punto de corte se establece en la puntuación en la que la probabilidad de ser satisfactorio es 0,50. Si los dos tipos de errores de clasificación no se consideran perjudiciales, se selecciona una puntuación de corte que reduzca el daño causado por los errores de clasificación (Perie y Thurlow, 2012).

1.2.1 Criterios para la selección del método

A pesar de la diversidad de métodos para establecer puntos de corte, no todos estos son apropiados para cualquier tipo de instrumento de evaluación. Tanto las características del instrumento y las restricciones adicionales, como el tiempo para desarrollar el taller de puntos de corte, la complejidad de la tarea cognitiva que se asigna a los jueces, entre otras, pueden reducir el número de métodos candidatos. Siguiendo a Aranguren y Hoszowski (2017), la elección de uno u otro método dependerá de, al menos, los siguientes criterios:

- Experiencia previa con el método.
- Facilidad de aplicación (en términos de tiempo, número de tareas y demanda cognitiva).
- Evidencia de validez del método.
- Heterogeneidad de los ítems utilizados en la evaluación.

Teniendo en cuenta estos cuatro criterios, la balanza se inclina a favor del método Bookmark para su uso en el Icfes por las siguientes razones:

a. Con respecto al primer criterio, el Icfes ha tenido experiencia previa con *Bookmark*, pues se han realizado diferentes talleres de establecimiento de estándares de desempeño desde el 2009, en exámenes como Saber 11.°, Saber Pro y Saber 3.°, 5.° y 9.°.

b. Con relación al segundo criterio, *Bookmark* no es holístico: en Bookmark, los jueces no tienen que revisar

ni establecer juicios a partir de las respuestas completas a varios ítems de una muestra de los evaluados. Por estas razones, *Bookmark* es menos costoso en términos de tiempo y tareas cognitivas que los métodos de tipo B y C. *Bookmark* tampoco se centra en juicios sobre las habilidades que exhiben una muestra de evaluados, por lo cual resulta menos costoso, en términos de tiempo, que los métodos de tipo D.

c. Aunque *Bookmark* y Angoff exigen juicios sobre la probabilidad de responder correctamente un ítem, en el primer caso, la probabilidad se mantiene constante, mientras que en Angoff los jueces tienen que estimar una probabilidad de respuesta que puede variar para cada ítem. Estos hechos generan que el ejercicio de estimación en Angoff sea más complejo cognitivamente que en *Bookmark*. Según un estudio realizado por el NAEP (National Assessment of Educational Progress), los hallazgos sugieren que Bookmark tiene la ventaja de ser menos costoso en términos de tiempo y tareas cognitivas (Peterson *et al*., 2011). Por tanto, en relación con el segundo criterio), *Bookmark* tendría ventajas sobre Angoff.

d. Con respecto al tercer criterio, los métodos *Bookmark* y Angoff tienen ventajas comparables y la percepción de su confiabilidad es alta, como lo confirma un estudio realizado por el NAEP (Peterson *et al*., 2011). De hecho, la confiabilidad y validez de estos dos métodos es similar. Sin embargo, no existen amplios estudios que comparen los diferentes tipos de métodos en términos de sus evidencias

de validez; de modo que el tercer criterio no permite señalar ventajas adicionales de *Bookmark* sobre Angoff.

e. Finalmente, frente al cuarto criterio, *Bookmark* da cuenta de los puntos de corte de pruebas en las que se incluyen tanto ítems de selección múltiple con única respuesta como ítems de respuesta construida. Ambos tipos de preguntas son desarrolladas para las evaluaciones que realiza el Icfes.

En conclusión, dados los criterios expuestos, *Bookmark* resulta ser el método más conveniente para el establecimiento de puntos de corte en las pruebas que realiza el Icfes.

1.2.2 Pasos para la aplicación de Bookmark

Insumos

El insumo principal para la ejecución de Bookmark es un cuadernillo con un conjunto de ítems (preguntas) de la prueba que están ordenados por la habilidad θ requerida para contestar correctamente al ítem dada una PR fija, la cual puede ser cualquier valor entre 0,5 y 0,8, siendo el más usado 0,67 o 2/3. Se recomienda que el *cuadernillo* esté compuesto mayoritariamente por ítems usados en una forma aplicada a la población evaluada (Cizek, 2012). Si existen diferencias notables en el valor θ requerido para dos ítems consecutivos en el cuadernillo, se incluirán ítems que minimicen estas diferencias. De este modo, se espera que los ítems del cuadernillo de ítems

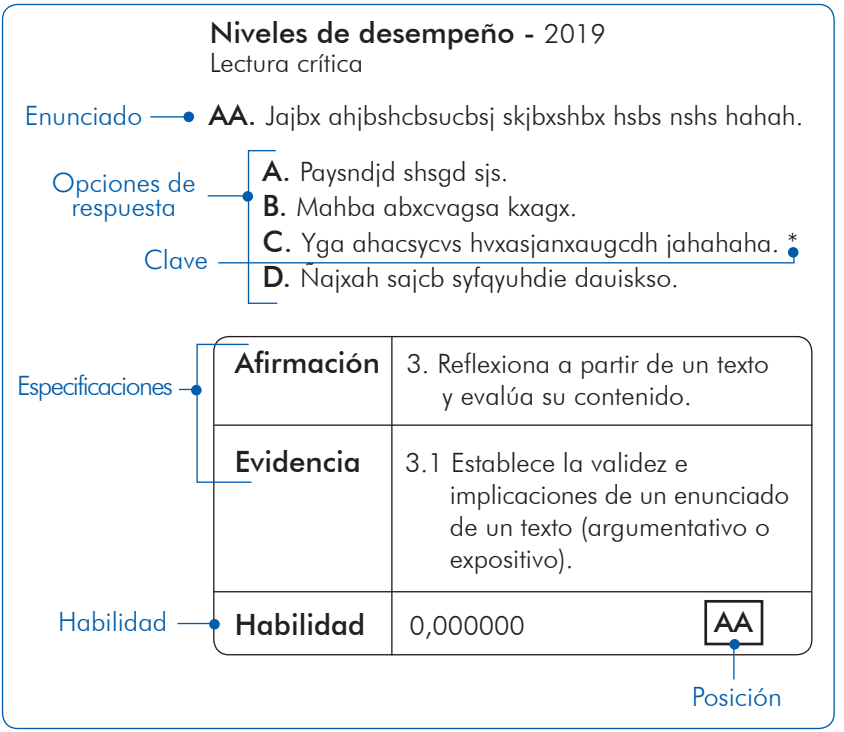
ordenado (CIO) representen las especificaciones de la prueba y estén uniformemente distribuidos a lo largo de la escala de habilidad (Aranguren y Hoszowski, 2017).

Aunque no hay una definición precisa del número de ítems que debe contener el cuadernillo ordenado, la cantidad final del CIO dependerá tanto del número de ítems que tenga la forma aplicada como de la cantidad que se agreguen para subsanar los vacíos en habilidad, los ejemplos en la literatura sugieren un CIO con alrededor de 50 ítems (Cizek, 2012; Lewis *et al*, 2012). Sobre la estructura del CIO, es importante aclarar que se debe presentar solo un ítem por hoja. En la esquina inferior derecha de cada hoja se muestra la ubicación del ítem (1 para el de menor dificultad, 2 para el siguiente, hasta llegar al ítem más difícil), usando una fuente que haga sobresalir el número, para que sea fácilmente identificable. Adicionalmente, se mostrará la información relativa a cada ítem, como la afirmación y la evidencia a la que corresponde, así como la información estadística del ítem: dificultad o habilidad o ambas. Cada ítem debe aparecer como fue administrado a los evaluados, excepto porque la opción correcta se señalará con un asterisco.

Al respecto de los contextos de los ítems, se recomienda colocarlos en un *cuadernillo complementario* en el caso en que este sea físico. Cada ítem del *cuadernillo* debe mostrar en un recuadro aparte en qué página del cuadernillo complementario se encuentra el contexto (Aranguren y Hoszowski, 2017). Si el cuadernillo es digital, es conveniente que en cada página

aparezca el contexto y el ítem con la opción correcta señalada[3]. Como insumo adicional, los jueces tendrán la descripción cualitativa de los niveles de desempeño y los formatos mostrados en el anexo 4.2. La estructura de una página del CIO se presenta en la ilustración 1.

Ilustración 1. *Estructura de página del CIO*



3 Ya sea en papel o digital, los panelistas contarán con un formato individual en el que podrán registrar sus marcas y hacer anotaciones sobre ítems (qué miden, por qué un determinado ítem es más fácil o difícil que otro, etc.).

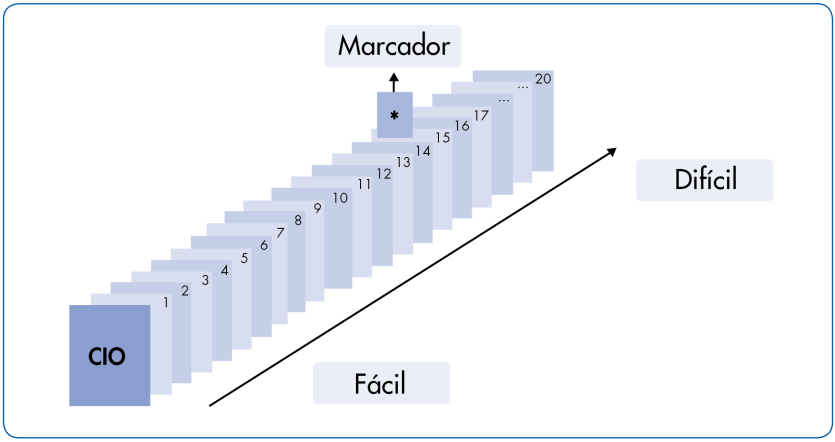
Descripción general de la tarea del método *Bookmark*

De acuerdo con Lewis *et al* (2012), se recomienda contar con un grupo de jueces que puede variar en número de integrantes (entre 10 y 24 jueces). Los jueces deben representar los diferentes niveles socioeconómicos de la población que atienden, géneros, tipo de institución (pública, privada) y minorías étnicas. En el caso del Icfes, en diferentes talleres de puntos de corte se ha contado con un panel de entre 10 y 12 jueces. El número de jueces puede variar a criterio de los diseñadores de prueba y las instituciones encargadas de definir políticas educativas, dependiendo de las condiciones logísticas, de recursos y procurando cumplir con la representación requerida.

La tarea de cada uno de los jueces, siguiendo el método *Bookmark*, consiste en fijar una marca en el cuadernillo para cada punto de corte. Esto significa que, para cuatro niveles de desempeño, se pondrán tres marcas. Cada marca se asigna al último ítem que un estudiante mínimamente competente en consideración pueda responder correctamente con probabilidad dada por la PR definida (0,5, 2/3, etc.). En el caso del ejemplo presentado en la tabla 2, en el que se definieron 4 Niveles, se tendría un estudiante mínimamente competente para el nivel 1, uno para el nivel 2 y uno para el nivel 3. Consecuentemente, los jueces determinarán tres puntos de corte. La asignación de las marcas debe hacerse revisando los ítems en el orden en que aparecen en el CIO. En ningún caso debe revisarse el cuadernillo en el orden inverso o en desorden.

Para enfocar a los jueces en la tarea de *Bookmark*, es importante recalcar que su labor en este punto consiste específicamente en determinar hasta qué ítem del CIO cada uno de los *estudiantes mínimamente competentes* podrá contestar correctamente con una probabilidad dada por la PR (Herrera Ortiz *et al*., 2009: 45). Por ejemplo, si una prueba tiene dos niveles de desempeño *Reprobado - Aprobado* y una PR de 2/3, los jueces deben pensar en un estudiante mínimamente competente para ser aprobado. Así, si un juez considera que este estudiante puede responder correctamente con una probabilidad de al menos 2/3 hasta el ítem en la posición 15 del ordenamiento, debe señalar una marca sobre este ítem en el CIO (*ver* ilustración 2); es decir, para este juez, los estudiantes que tengan una habilidad mayor o igual a la requerida para responder el ítem en la posición 15, son las que quedarán clasificados en el nivel **aprobado**. La manera de definir el punto de corte definitivo se explica más adelante.

Ilustración 2. *Marcador sobre un CIO con 20 ítems*



El procedimiento total de la aplicación de *Bookmark* cuenta con cuatro fases (Lin, 2006). La primera es la capacitación de los jueces, y las demás corresponden a tres rondas de aplicación para establecer los puntos de corte. A través de cada ronda se intenta incrementar el consenso y reducir las diferencias entre los jueces, si existen. Las cuatro fases se especifican a continuación.

Fase 1: capacitación del panel de jueces

Durante la capacitación se presenta la DND, se realiza una breve presentación del modelo TRI que se aplica en la calificación de la prueba (específicamente, se explica en términos generales el concepto de curva característica), se enfatiza la importancia del establecimiento de puntos de corte y se inicia la capacitación en la metodología *Bookmark* (Lin, 2006).

Aranguren y Hoszowski (2017) sugieren que, para que los jueces entiendan el juicio de probabilidad que deben realizar, se les plantee la siguiente situación: "piense en un grupo representativo de estudiantes que se encuentren en el límite superior del nivel bajo. Dos de cada tres de estos estudiantes, ¿contestarían este ítem de manera correcta?". Adicionalmente, dentro de cada grupo se recomienda nombrar un líder, al cual se le instruye para que sirva de facilitador en la discusión, mantenga al grupo enfocado en la tarea y verifique el cumplimiento de los tiempos de la agenda (Lewis *et al*., 2012). El facilitador puede ser desempeñado por un profesional del área de la prueba quien puede moderar la discusión, pero en ningún caso tomar partido hacia una propuesta específica en la ubicación de una marca.

Fase 2: ronda 1

Los objetivos de esta ronda son: familiarizar a los jueces con el cuadernillo de ítems ordenados, colocar las primeras marcas y empezar a discutir la ubicación de estas (Lin, 2006). En esta ronda, los panelistas trabajan individualmente. Cada juez marca el ítem que, según su juicio, define cada punto de corte; para ello, los jueces cuentan tanto con la descripción específica de niveles como con un **formato individual para establecer marcas** (*véase* Anexo 3.2, formato 1) en el que dejan el registro en el espacio correspondiente a la primera ronda.

Con el propósito de facilitar la tarea de los jueces, la descripción específica de niveles obtenida en el taller de DND se puede usar tanto para concebir los estudiantes mínimamente competentes como para resolver la **pregunta central del método** *Bookmark*: ¿hasta qué ítem del ordenamiento puede responder correctamente un estudiante mínimamente competente con una probabilidad dada por la PR? Una manera razonable de usar la descripción de niveles para estos dos propósitos se expone en lo que sigue:

a. Para tener una idea clara de lo que es un estudiante mínimamente competente que está en el nivel *N*, este se define como aquel que cumple con las descripciones que hacen parte del nivel *N*, en un grado mínimo de apropiación y no cumple todas las descripciones del nivel *N*+1.

b. Para identificar el último ítem que el estudiante mínimamente competente puede responder correctamente, por ejemplo, con 2/3 de probabilidad, se sugiere lo siguiente: al iniciar la revisión de los ítems, siempre empezando por el primero del CIO, se debe reconocer un *salto cualitativo* entre las habilidades requeridas para responder dos ítems consecutivos en el cuadernillo. Así, si se quiere ubicar la primera marca y se encuentra un primer *salto cualitativo* entre los ítems 2 y 3 del CIO, la primera marca se pondrá en el ítem 2. Para ilustrar lo que es un salto cualitativo, considere el siguiente ejemplo: suponga que se quiere establecer la marca que separa a los estudiantes en nivel 1 de aquellos en nivel 2. Además, considere que los primeros seis ítems del ordenamiento solo requieren los CHD consignados en los descriptores del nivel 1. Si el ítem número siete del cuadernillo exige CHD asociados a los niveles 2 o 3, entonces hay un salto cualitativo entre el ítem seis y el ítem siete del cuadernillo: un estudiante mínimamente competente del nivel 1 tendrá los CHD para responder hasta el ítem 6, pero ya no tendrá los CHD necesarios para responder el ítem 7, pues este exige habilidades que tiene, como mínimo, un estudiante en el nivel 2. Como existe un *salto cualitativo* entre los ítems seis y siete, la marca se pondrá en el ítem 6. Para poner la siguiente marca, se ubica un salto cualitativo posterior en el ordenamiento de ítems.

Esta forma de usar la descripción cualitativa de niveles de desempeño es solo una estrategia para alimentar la toma de decisiones sobre la ubicación de las dos primeras marcas

en el cuadernillo y en ningún caso es una estrategia de uso obligatorio. Cada juez tiene la autonomía de proceder a indicar sus marcas de acuerdo con otros procesos de razonamiento, siempre y cuando estos respeten las fases del método *Bookmark* y respondan la pregunta central de este método.

Fase 3: ronda 2

En la segunda ronda, los jueces se reúnen en dos grupos. En cada grupo, los jueces deben ubicar, en el *formato grupal para establecer marcadores* (*véase* Anexo 3.2, formato 2), las marcas que sus compañeros de grupo establecieron en la primera ronda. La discusión en esta ronda se centrará en el primer y el último ítem que componen cada una de las tres marcas que se deben establecer. De esta forma, si las marcas para trazar el primer punto de corte de un grupo de cinco jueces corresponden a los ítems en la posición 6, 10, 12, 13 y 20 del CIO, se discutirán las razones para ubicar las marcas en los ítems 6 y 20. Una discusión similar se dará para las marcas correspondientes a los otros dos puntos de corte. Al finalizar esta discusión, cada juez debe, de forma independiente, reubicar sus marcas y dejar la evidencia en el formato grupal (Lin, 2006). En este punto, los jueces pueden recoger sus observaciones acerca de ítems particulares del cuadernillo, acerca la DND o acerca de la asignación de los grupos de evidencias a cada nivel. Estas observaciones alimentarán la discusión inicial que se dará en la ronda 3.

Fase 4: ronda 3

En esta ronda se reúne el grupo total de jueces y discuten sus marcas finales de la ronda 2, así como las observaciones sobre los ítems del ordenamiento y los descriptores de niveles. Luego de esta discusión, los jueces asignan una nueva marca en la columna de marca inicial del *formato 3* de puntos de corte (*véase* Anexo 3.2, formato 3). Siguiendo a Lin (2006), a partir de estas marcas iniciales de la ronda 3, expertos en psicometría o estadística informa a los jueces sobre la distribución de la población, dadas estas marcas (el porcentaje de evaluados que quedaría ubicado en cada nivel). Aunque esta información puede ser un insumo para que los jueces generen expectativas realistas sobre el desempeño de los estudiantes, es importante señalar a los jueces que la toma de decisiones sobre las posiciones de las marcas debe estar motivada, fundamentalmente, por una discusión en la que se evalúen las observaciones de los jueces sobre ítems particulares, los descriptores, la asignación de descriptores a niveles y el conocimiento de la población.

Después de observar la distribución de la población dadas sus marcas, el grupo total de jueces señala, nuevamente de manera individual, las tres marcas. Estas marcas finales quedan registradas en la columna de marcas finales del *formato 3*. Los puntos de corte se extraerán de la siguiente manera (Cizek 2012, Lane *et al*., 2016): cada punto de corte corresponde a la media o mediana (se elige el valor que mejor represente al conjunto de datos) de las habilidades requeridas para responder correctamente los ítems marcados por los

jueces con una PR fija, para dicho punto de corte. Así, por ejemplo, si los jueces asignaron la primera marca a los ítems en la posición 3, 4, 6 y 7, las habilidades asociadas a cada uno de estos ítems son promediadas (o su mediana calculada). Dicha habilidad media (o mediana) corresponderá al primer punto de corte. Lo mismo se aplica para la obtención de los demás puntos de corte. El puntaje mínimo para alcanzar cada nivel será obtenido a partir del punto de corte a través de una transformación a la escala de puntaje de la prueba.

El taller de puntos de corte finaliza, primero, con la revisión del ordenamiento de evidencias del taller de DND (para ajustarlo a la luz de las observaciones realizadas en los dos talleres). Segundo, se concluye con la evaluación, por parte de los jueces, del desarrollo de ambos talleres y sus sugerencias de mejora. Para dicha evaluación, los jueces diligenciarán un formato de encuesta (*véase* Anexo 3.2, formato 4), en el cual se encuentra un ejemplo de agenda de tres días para la realización de los dos talleres, el de definición de niveles de desempeño y el de puntos de corte, que puede servir de guía para planear y organizar los procedimientos descritos en este documento.

2. REFERENCIAS

Aranguren, M. y Hoszowski, A. (2017). Aprender 2016. Serie de documentos técnicos 5: Bookmark, Establecimiento de puntos de corte. Ministerio de Educación y Deportes: Buenos Aires. Disponible en http://www.educacion.gob.ar/data_storage/file/documents/manual-bookmark595bd361cf4e7.pdf

Bejar, I. (2008). Standard Setting: What is it? Why is it impotant? [Versión electrónica] *R & D. Connections*, 7, 1 - 6. Extraído el 6 Mayo, 2020 de: https://www.ets.org/Media/Research/pdf/RD_Connections7.pdf

Bloom, B. (Ed.) (1956) Taxonomy of educational objectives: The classification of educational goals: Handbook I, cognitive domain. New York, Toronto: Longmans, Green.

Cizek, G. (2011). An Introduction to Contemporary Standard Setting: Concepts, Characteristics, and Contexts En: Gregory J. Cizek (ed.) (2011). *Setting Performance Standards Foundations, Methods, and Innovations*. Routledge. NY.

Cizek, G. J. (Ed.). (2012). *Setting performance standards: Foundations, methods, and innovations*. Routledge. New York, NY: Routledge.

Cizek, G. J. (2012b). *An Introduction to Contemporary Standard Setting*. Cizek, G. J. (Ed.). (2012). *Setting performance standards: Foundations, methods, and innovations*. New York, NY: Routledge.

Cizek, G. y Buch, M. (2007). Standard setting: A guide to establishing and evaluating performance standards for tests. Sage Publications. Londres.

Egan, K., Schneider, C. y Ferrara, S. (2012) "Performance Level Descriptors". En: Setting Performance Standards. Second Edition (Cizek, G. Ed.). Routledge. NY: 79-106.

Herrera Ortiz, M., Monroy Cazorla, L., y Benavides Posadas, D. (2009). Establecimiento de estándares en un examen criterial. *Cuaderno Técnico*, 3.

Huynh, H. (2000). On item mapping and statistical rules for selecting binary items for criterion-referenced interpretation and bookmark standard setting. Paper presented at the annual meeting of National Council on Measurement in Education, New Orleans.

Huynh, H. (2006). A Clarification on the Response Probability Criterion RP67 for Standard Settings Based on Bookmark and Item Mapping. Educational Measurement: Issues and Practice. Vol. 25, Issue 2.

Kingston, N. M. y Tiemann, G. C. (2012). *Setting Performance Standards on Complex Assessments* (pp 220-221). New York, NY: Routledge.

Lane, M. Raymondand T. y. Haladyna (2016). Handbook of Test Development. Second Edition. Routledge. NY.

Lewis, D. M., Mitzel, H. C., Mercado, R. L., y 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.

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.

Mislevy, R. et al (2003). *A brief introduction to evidence-centered design*. Educational Testing Service, Princeton, NJ. July 2003

Mislevy, R. et al (2017). *Assessing Model-Based Reasoning using Evidence- Centered Design: A Suite of Research-Based Design Patterns*. Springer.

Perie, M. y Thurlow, M. (2012). *Setting Achievement Standards on Assessments for Students with Disabilities*. Cizek, G. J. (Ed.). (2012). *Setting performance standards: Foundations, methods, and innovations* (pp 372-373). New York, NY: Routledge.

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

Peterson CH, Schulz EM & Engelhard Jr G (2011). Reliability and validity of bookmark□based methods for standard setting: comparisons to Angoff-based methods in the National Assessment of Educational Progress. Educ Meas Issues Pract. 2011; 30(2): 3–14.

Plake, B., Huff, K. y Reshetar, R. (2010). Evidence-Centered Assessment Design as a Foundation for Achievement-Level Descriptor Development and for Standard Setting. Applied Measurement in Education, 23: 342–357.

Reckase, M. y Jing Chen (2012). *The Role, Format, and Impact of Feedback to Standard Setting Panelists*. En: Cizek, G. J. (Ed.). (2012). *Setting performance standards: Foundations, methods, and innovations*. Routledge. New York, NY: Routledge

Zieky, M. y Perie, M. (2006). A Primer on Setting Cut Scores on Tests of Educational Achievement. Educational Testing Service.

Zieky, M., Perie, M., y Livingston, S. (2008). Cut scores: A manual for setting performance standards on educational and occupational tests. Princeton, NJ: Educational Testing Service.

Zieky, M. J. (2012). So Much Has Changed. An Historical Overview of Setting Cut Scores. En: Cizek, G. J. (Ed.). (2012). *Setting performance standards: Foundations, methods, and innovations*. Routledge. New York, NY: Routledge

3. ANEXOS

3.1 Taxonomía de Bloom

Ilustración 1. *Estructura de página del CIO*

TAXONOMÍA DE BLOOM DE HABILIDADES DE PENSAMIENTO (1956)

CATEGORÍA	CONOCIMIENTO RECOGER INFORMACIÓN	COMPRENSIÓN CONFIRMACIÓN APLICACION	APLICACIÓN HACER USO DEL CONOCIMIENTO	ANÁLISIS (ORDEN SUPERIOR) DIVIDIR, DESGLOSAR	SINTETIZAR (ÓRDEN SUPERIOR), REUNIR, INCORPORAR	EVALUAR (ÓDEN SUPERIOR) JUZGAR EL RESULTADO
Descripción Las habilidades que se deben demostrar en este nivel son:	Observación y recordación de información; conocimiento de fechas, eventos, lugares; conocimiento de las ideas principales; dominio de la materia.	Entender la información; captar el significado; trasladar el conocimiento a nuevos contextos; interpretar hechos; comparar, contrastar; ordenar, agrupar; inferir las causas predecir las consecuencias.	Hacer uso de la información; utilizar métodos, conceptos, teorías, en situaciones nuevas; solucionar problemas usando habilidades o conocimientos.	Encontrar patrones; organizar las partes; reconocer significados ocultos; identificar componentes.	Utilizar ideas viejas para crear otras nuevas; generalizar a partir de datos suministrados; relacionar conocimiento de áreas diversas; predecir conclusiones derivadas.	Comparar y discriminar entre ideas; dar valor a la presentación de teorías; escoger basándose en argumentos razonados; verificar el valor de la evidencia; reconocer la subjetividad.
Que Hace el Estudiante	El estudiante recuerda y reconoce información e ideas además de principios aproximadamente en misma forma en que los aprendió.	El estudiante esclarece, comprende, o interpreta información en base a conocimiento previo.	El estudiante selecciona, transfiere, y utiliza datos y principios para completar una tarea o solucionar un problema.	El estudiante diferencia, clasifica, y relaciona las conjeturas, hipótesis, evidencias, o estructuras de una pregunta o aseveración.	El estudiante genera, integra y combina ideas en un producto, plan o propuesta nuevos para él o ella.	El estudiante valora, evalúa o critica en base a estándares y criterios específicos.
Ejemplos de Palabras Indicadoras	- define - lista - rotula - nombra - identifica - repite - quién - qué - cuando - donde - cuenta - describe - recoge - examina - tabula - cita	- predice - asocia - estima - diferencia - extiende - resume - describe - interpreta - discute - extiende - contrasta - distingue - explica - parafrasea - ilustra - compara	- aplica - demuestra - completa - ilustra - muestra - examina - modifica - relata - cambia - clasifica - experimenta - descubre - usa - computa - resuelve - construye - calcula	- separa - ordena - explica - conecta - divide - compara - selecciona - explica - infiere - arregla - clasifica - analiza - categoriza - compara - contrasta - separa	- combina - integra - reordena - substituye - planea - crea - diseña - inventa - que pasa si? - prepara - generaliza - compone - modifica - diseña -plantea hipótesis - inventa - desarrolla - formula - reescribe	- decide - establece gradación - prueba - mide - recomienda - juzga - explica - compara - suma - valora - critica - justifica - discrimina - apoya - convence - concluye - selecciona - establece rangos - predice - argumenta

Tomado de:
http://eduteka.icesi.edu.co/pdfdir/TaxonomiaBloomCuadro.pdf

3.2 Formatos

Formato 1: formato individual para establecer marcadores.

En este formato los jueces consignan cada una de sus marcas en la revisión inicial del cuadernillo de ítems ordenados.

Nombre del juez:			
Prueba:			
Método *Bookmark* individual			
Ronda	**Primer Marca**	**Segunda Marca**	**Tercer Marca**
1			
Observaciones:			

Formato 2: formato grupal para establecer marcadores.

En este formato cada miembro del grupo colocará sus marcas y las de sus compañeros, al inicio y al final de la ronda 2.

Prueba:						
Método *Bookmark* Ronda 2						
Nombre del juez:	**Primer Marca**		**Segunda Marca**		**Tercer Marca**	
	Inicio	Final	Inicio	Final	Inicio	Final
1.						
2.						
3.						
4.						
5.						
Observaciones:						

Formato 3: formato general de marcadores.

En este formato, un moderador del grupo, registrará las marcas finales de la ronda 3 de la totalidad de los jueces.

Prueba:						
Método *Bookmark* Ronda 3						
Nombre del juez:	**Primer Marca**		**Segunda Marca**		**Tercer Marca**	
	Inicio	Final	Inicio	Final	Inicio	Final
1.						
2.						
3.						
4.						
5.						
6.						
7.						
8.						
9.						
10.						
Observaciones:						

Formato 4: evaluación del desarrollo de los talleres.

Nombre del juez:

Al respecto del trabajo realizado durante estos dos días, marque con una x la casilla correspondiente, su acuerdo o desacuerdo con las siguientes afirmaciones:

Taller de Descripción de Niveles de Desempeño

Afirmación	Sí	No
Fue clara y suficiente la información dada sobre el Diseño Centrado en Evidencias.		
Fue clara y suficiente la información dada sobre las especificaciones de la prueba.		
Fue claro el objetivo del taller.		
Fue suficiente el tiempo de discusión para la definición de los descriptores de los niveles de desempeño.		

Taller de Establecimiento de Puntos de Corte

Afirmación	Sí	No
Fue clara y suficiente la información dada sobre el módelo 3PL.		
Fue clara y suficiente la información dada sobre el método *bookmark*.		
Entendió qué era un *estudiante frontera*.		

Continúa

Afirmación	Sí	No
Fueron comprensibles los ítems en el cuadernillo ordenado (ronda 1).		
Fue suficiente el tiempo de la ronda 1 para establecer los primeros *bookmarks*.		
Fue suficiente el tiempo de la ronda 2 para debatir al interior de los grupos.		
Fueron fáciles de usar los formatos y materiales usados en la ronda 2.		
Fue suficiente el tiempo de la ronda 3 para debatir los bookmarks de todos los jueces.		
El proceso para establecer los puntos de corte fue justo y careció de sesgos.		

Comentarios adicionales y sugerencias de mejora:

3.3 Agenda para el establecimiento de estándares de desempeño

Taller de Establecimiento de Estándares: niveles de desempeño y puntos de corte.

El taller consta de dos partes complementarias: el diseño de los Descriptores de Niveles de Desempeño (DND) y el establecimiento de puntos de corte (PC). DND se desarrolla durante el primer día y PC se desarrolla durante el segundo y el tercer día:

1. Descriptores de Niveles. Este día se divide en 3 momentos:
   - 1.1 Presentación de los objetivos e insumos del taller de DND:
     - 1.1.1 Breve descripción de qué es un Nivel de Desempeño.
     - 1.1.2 Presentación de los facilitadores y coordinadores.
     - 1.1.3 Explicación del DCE.
     - 1.1.4 Exposición de especificaciones de prueba: afirmaciones y evidencias.
   - 1.2 La ordenación de las evidencias.
     - 1.2.1 Ordenación de evidencias.
     - 1.2.2 Ajuste del continuo
   - 1.3 Asignación de los descriptores a los 4 niveles de desempeño.

2. Puntos de corte
   - 2.1 Presentación: objetivo del taller de PC.
     - 2.1.1 Conceptos básicos de TRI.
   - 2.2 Método Bookmark.
     - 2.2.1 Ronda 1.
     - 2.2.2 Ronda 2.
     - 2.2.3 Ronda 3.

3. Revisión del ordenamiento de evidencias, evaluación del taller y cierre.

Día 1

Hora	Actividad	Responsable
8:00 – 10:00	Presentación, objetivo del taller y sensibilización.	Gestor de pruebas
	Diseño Centrado en Evidencias.	
	Especificaciones de prueba.	
10:00 – 10:15	Receso.	Todos
10:15 – 12:00	Ordenación de evidencias I: en este espacio trabajan en dos grupos de cinco personas.	Jueces
12:00 – 13:00	Almuerzo.	Todos
13:00 – 15:00	Ordenación de evidencias II: debate entre los 10 jueces.	- Gestor. - Facilitador. - Jueces.
15:00 – 15:15	Receso.	- Gestor. - Facilitador. - Jueces.
15:15 – 17:00	Asignación de Niveles a Descriptores.	- Gestor. - Facilitador. - Jueces.

Día 2

Hora	Actividad	Responsable
8:00 – 9:30	Presentación.	- Gestor de prueba. - Estadístico. - Facilitador.
	TRI (3PL).	
	Método *Bookmark* (conceptos centrales, rondas y materiales de trabajo).	
	Ejemplo de asignación de marcas.	
9:30 – 12:00	Ronda 1: trabajo individual en banco (tendrán un cuadernillo en PDF con posición del ítem, contexto e ítem, clave, afirmación, evidencia y un formato en papel en el que anotarán las marcas).	Jueces
12:00 – 13:00	Almuerzo.	Todos
13:00 –14:30	Ronda 2: trabajo por subgrupos de 5 miembros. Debate centrado en las marcas puestas en la ronda 1 (se trabajará en dos salas en las que se proyectarán los ítems que sean objeto del debate).	- Gestor de prueba. - Estadístico. - Psicometría. - Asesores de prueba. - Jueces.
14:30 – 14:45	Receso.	Todos

Continúa

Hora	Actividad	Responsable
14:45 – 17:00	Ronda 3: Comparación de puntos con todo el grupo. Discusión de ítems que generen dudas. Luego, presentación % de evaluados que quedarían ubicados en cada nivel, de acuerdo con los puntos de corte propuestos por los jueces al final de la ronda 2. Posteriormente, los 10 jueces discuten para llegar a acuerdos y definir las marcas finales.	- Gestor de prueba. - Estadístico. - Psicometría. - Asesores de prueba. - Jueces.

Nota: Es indispensable que en la ronda 1 cada juez cuente con un computador, si el cuadernillo se presenta en versión digital.

Día 3

Hora	Actividad	Responsable
8:00 – 9:30	Revisión del ordenamiento de ítems.	Jueces
	Revisión de la asignación de descriptores a niveles de desempeño.	
9:30 – 9:45	Receso.	Todos
9:45 – 10:30	Evaluación del desarrollo de los Talleres.	Jueces
10:30 – 11:30	Cierre y entrega de certificados de participación a los Jueces.	- Gestor. - Facilitador.

Calle 26 N.° 69-76, Torre 2, Piso 15, Edificio Elemento, Bogotá, D. C., Colombia • www.icfes.gov.co
Líneas de atención al usuario: Bogotá Tel.: (57+1) 484-1460 PBX: (57+1) 484-1410 -
Gratuita nacional: 018000-519535