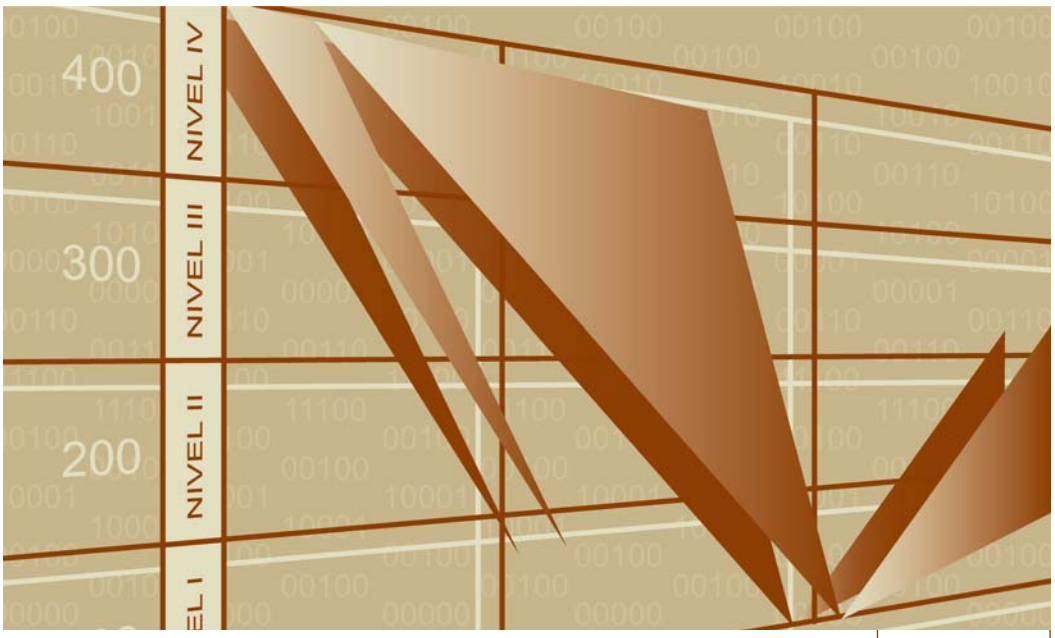
Instituto Nacional para la Evaluación de la Educación

Manual técnico

Establecimiento de niveles de competencia



13

Excale
Exámenes de la Calidad y el Logro Educativos

Dirección de Pruebas y Medición

El *Manual técnico para el Establecimiento de niveles de competencia* es un documento de uso oficial elaborado por la Dirección de Pruebas y Medición del Instituto Nacional para la Evaluación de la Educación.

DISEÑO Y DIAGRAMACIÓN:
Karla Sandra Ramírez Quintero

ILUSTRACIÓN DE PORTADA:
Pablo Josué Pulido Ramírez

D. R. © 2006 Instituto Nacional para la Evaluación de la Educación
José Ma. Velasco 101, Col. San José Insurgentes
México, D. F.

Primera edición: noviembre de 2006

Impreso en México

INSTITUTO NACIONAL PARA LA EVALUACIÓN DE LA EDUCACIÓN

Dirección de Pruebas y Medición

Manual técnico

Establecimiento de niveles de competencia

REDACTADO POR:

Jesús Jornet Meliá
Eduardo Backhoff Escudero

ÍNDICE

INTRODUCCIÓN

La determinación de estándares para la interpretación de los niveles de logro de los Exámenes para la Calidad y el Logro Educativos (Excale) constituye un proceso complejo en la construcción de las pruebas, el cual implica elementos de validación. La estrategia y los procedimientos utilizados para este propósito deben basarse en métodos sustentados en la experiencia que las diversas agencias e instituciones de evaluación, nacionales e internacionales han desarrollado y, a partir de ella, crear las interpretaciones de pruebas con sustento en estándares cuidadosamente diseñados, mediante soluciones ajustadas específicamente a los requerimientos de diseño, construcción y aplicación de los Excale.

Tal como lo señala el estándar 4.19 de la APA, AERA y NCME (1999),[1] una de las claves de validez de los estándares de interpretación de los niveles de logro es la transparencia del proceso llevado a cabo para su determinación; la información recabada en este tipo de procesos debe ser la primera garantía acerca de la calidad de los niveles de logro identificados. Dentro del Plan maestro de Desarrollo de los Excale, la determinación de los niveles de logro representa un punto culminante del diseño de los instrumentos, pues aporta el sistema que interpreta los resultados de la evaluación.

El presente manual se propone detallar el proceso para establecer los niveles de logro. Determinar el sistema de interpretación de las pruebas es un aspecto delicado, pues supone definir los criterios en virtud de los cuales podrá valorarse la calidad del aprendizaje.

Los Excale han sido diseñados como pruebas apegadas al currículo mexicano y, aunque a partir de los puntajes de habilidad total, se puede graduar y ordenar la muestra de alumnos evaluados en función de su nivel de logro en las pruebas, esto no constituye por sí mismo una interpretación acerca de si el logro que se observa es insuficiente, suficiente o elevado respecto al dominio del currículo; para poder interpretar en este sentido, se requiere establecer un proceso en el cual especialistas en educación (técnicos en currículo, investigadores educativos, autores de libros de texto y profesores) aporten los criterios de evaluación. De este modo, es necesario dotar a los Excale de un componente criterial, tal como se especifica en la definición del tipo de pruebas que el INEE requiere para cumplir sus objetivos.

En este manual se describe la información relativa al método para la determinación de estándares, así como a su establecimiento e interpretación.

[1] "Cuando las interpretaciones propuestas consideran uno o más puntos de corte, la lógica y los procedimientos usados para establecer los puntos de corte deben estar claramente documentados." (p. 59).

1 PROCEDIMIENTO DE ELECCIÓN DEL METÓDO

La cultura de evaluación que ha conformado el INEE a lo largo de su existencia tiene un común denominador: el trabajo colaborativo como base de diseño y desarrollo de instrumentos, así como para la realización de los diversos planes de evaluación emanados de ello. Subyacente a este planteamiento, se identifican:

- el concepto de evaluación como un proceso multidisciplinario de especialistas e investigadores en educación, profesores en activo y técnicos en metodología de evaluación, y
- la complementariedad metodológica entre recursos cuantitativos y cualitativos, como el medio más fiable y válido, acorde con el objeto de evaluación.

Para la elección del método se desarrolla un seminario, en el cual se revisan aspectos tales como el concepto "estándar", las consideraciones acerca de los métodos para determinar estándares y puntos de corte, además del concepto de validez de los estándares, las estrategias para su validación, los criterios de calidad y el establecimiento del proceso para determinar los niveles de logro.

El formato de trabajo de este seminario se fundamenta en dos tipos de actividades:

1. Expositiva, por parte del conductor del seminario, y
2. debate de propuestas

Los productos del seminario podemos resumirlos en:

1. Elección del modelo de determinación de estándares, como una adaptación del método Bookmark (o "del marcador")
2. Definición de las categorías y etiquetas del INEE para los niveles de logro
3. Formación inicial de los coordinadores de pruebas como coordinadores de los comités de establecimiento de estándares e identificación de puntuaciones de corte.

MODELO DEL INEE PARA LA DETERMINACIÓN DE NIVELES DE LOGRO DE LAS PRUEBAS Excale

2

El rendimiento, evaluado a partir de las pruebas Excale, es un constructo de contenido curricular y se asume de carácter continuo. El supuesto es que el Excale actúa como un instrumento que evalúa personas con respecto a su nivel de dominio curricular, y esa evaluación es representativa de la graduación del dominio de contenidos. Por ello, la determinación de niveles diferenciales de logro que simplifiquen y faciliten la interpretación de los resultados de la prueba se basa en la identificación de puntuaciones de corte que indiquen, con una elevada fiabilidad, tipologías diferenciales de ejecución a lo largo de dicho continuo.

Si bien no puede asumirse que haya un procedimiento para la determinación de estándares y puntos de corte que sobresalga sobre los demás por su calidad, pertinencia o la bondad de los resultados, sí parecen claros algunos principios básicos que ponen de manifiesto las ventajas de los métodos centrados en los reactivos de carácter mixto (juicio-empírico) y con retroalimentación de información a los participantes en los comités de desarrollo. Asimismo, creemos que hay que asumir la arbitrariedad que preside toda interpretación de resultados psicopedagógicos y sociales —en el sentido de su elegibilidad, y no que sean caprichosa—. En este sentido, el procedimiento para la determinación de niveles de logro debe cumplir básicamente con los siguientes objetivos:

1. Que el sistema de interpretación sea representativo de las opiniones que los expertos en educación tienen acerca de lo que puede dar como resultado el sistema educativo en cada una de las materias evaluadas, por lo que este debe basarse en procesos de consenso intersubjetivo —de expertos en cada una de ellas—, debidamente dirigidos y evaluados;

2. que considere la implementación real del sistema educativo a través de una población tan diversa como es la de México, por lo que debe atenderse la realidad distribucional y las características diferenciales del comportamiento de los alumnos mexicanos ante los Excale; y

3. que permita validar los niveles resultantes a partir de estudios empíricos.

La calidad de los estándares de interpretación de los niveles de logro debe ser tal que permita su interpretación para la evaluación actual, así como facilitar interpretaciones de carácter longitudinal y transversal de la situación y evolución de los resultados del sistema educativo.

Para ello, el proceso de determinación de los niveles de logro debe basarse en el trabajo de dos comités: de especialistas en currículo y en investigación educativa; y de profesores(as), denominados en lo sucesivo comités 1 y 2, respectivamente; el primero tiene como finalidad la determinación de las competencias, habilidades y contenidos característicos de cada nivel de logro, mientras que la tarea del segundo es la identificación de las puntuaciones de corte correspondientes en la prueba.

2.1 Determinación del sistema de interpretación: etiquetas y descriptores

Previamente al proceso de determinación del sistema de interpretación, el INEE estableció la categorización a utilizar para la interpretación de las pruebas.

La identificación de niveles se basa habitualmente en un sistema de tres a seis categorías, y cada una suele estar representada por una etiqueta representativa del nivel de logro en la materia. La finalidad del sistema de etiquetas es disponer de una referencia de determinación del tipo de dominio que poseen estudiantes característicos de cada nivel de logro, de modo que los comités que trabajen en la identificación de los contenidos, competencias y habilidades

correspondientes a cada nivel, puedan disponer de un marco de interpretación común para todos los Excale. El sistema de categorías, por otra parte, también sirve como base de síntesis para la comunicación de los mismos.

Los criterios para identificar el sistema de etiquetas son los siguientes:

1. Simplicidad, suficiente como para:
   - sintetizar la información en un reducido número de categorías
   - identificar las categorías con etiquetas fácilmente comprensibles por las diferentes audiencias a que se dirige la evaluación.
2. Valor diferencial de las categorías. No obstante la simplicidad del sistema, éste debe permitir discriminar de forma suficiente entre tipos habituales de alumnos, desde los que no llegan a poseer un dominio suficiente para avanzar en el aprendizaje de la materia, hasta aquellos que llegan a mostrar un logro muy elevado.
3. Las etiquetas, si bien representan niveles de logro, deben ser entendidas como meros identificadores, y a ser posible, de carácter ecléctico, de forma que no puedan ser interpretadas de forma negativa por la población (razón por la cual deben evitarse términos como, por ejemplo "inferior", "reprobado", "superior" o equivalentes).
4. Deben evitarse etiquetas ambiguas o que incluyan tecnicismos de difícil comprensión para la sociedad en general, que en definitiva será la receptora final del informe de resultados de la evaluación.
5. En cualquier caso, tanto la categorización propuesta como las etiquetas elegidas, deben ser susceptibles de revisión a partir de:
   - el funcionamiento de la prueba, pues hay que considerar si la prueba dispone de capacidad suficiente de discriminación para el sistema previsto
   - la opinión de los diversos comités implicados en la determinación de estándares de interpretación e identificación de puntuaciones de corte.

Así, la Dirección de Pruebas y Medición del INEE, apoyada por los participantes en el seminario inicial, y considerando los criterios mencionados, establece la categorización y definición de etiquetas. Resultado de los procesos anteriores, las etiquetas establecidas se muestran en la Tabla 1.

Tabla 1. Categorías y definición base de etiquetas para los Excale

- **Por debajo del nivel básico:** indica carencias importantes en el dominio curricular de los conocimientos, habilidades y destrezas escolares que expresan una limitación para poder seguir progresando satisfactoriamente en la materia.
- **Básico:** indica el dominio imprescindible suficiente, mínimo, esencial, fundamental, o elemental de conocimientos, habilidades y destrezas escolares necesarias para poder seguir progresando satisfactoriamente en la materia.
- **Medio:** indica un dominio sustancial (adecuado, apropiado, correcto o considerable) de conocimientos, habilidades y destrezas escolares, que pone de manifiesto un buen aprovechamiento de lo previsto en el currículo.
- **Avanzado:** indica un dominio muy elevado (intenso, inmejorable, óptimo o superior) de conocimientos, habilidades y destrezas escolares que refleja el aprovechamiento máximo de lo previsto en el currículo.

2.2 Determinación de los niveles de logro

El primer comité, que denominamos Comité de Niveles de Logro (CNL), se encarga de la elaboración de descriptores, mientras que el segundo, que denominamos Comité de identificación de Puntuaciones de Corte (CPC), se encarga de identificar los reactivos que servían de punto de inflexión entre dos niveles de logro, los cuales dirigen la identificación de las puntuaciones de corte que separan los niveles de logro.

La síntesis de las fases de trabajo a desarrollar para la determinación de los niveles de logro de los Excale se recoge en la Tabla 2.

Tabla 2. Modelo del INEE para la determinación de estándares o niveles de logro (NL)

Responsable	Forma de Trabajo	Productos
Fase 1. Elección del Modelo para la determinación de Niveles de Logro (NL) del Excale		
Seminario • Conductor • Consejo Técnico del INEE • Dirección General de Pruebas del INEE [2] • Representantes de la SEP	**Seminario** 1.Formación 2.Debate de propuestas 3.Decisiones. Elección del modelo del INEE	• Elección del modelo de determinación de estándares • Recomendaciones para la definición de categorías y etiquetas • Formación inicial de los coordinadores de pruebas
Fase 2. Determinación del sistema de interpretación: etiquetas y descriptores		
Dirección de pruebas del INEE		Categorización y definición de etiquetas
Fase 3. Determinación de Niveles de Logro (1/2)		
Momento 0. Elaboración de las Tablas de Especificaciones		
Dirección de pruebas del INEE		Propuesta de clasificación de especificaciones de subdominios, ordenados por dificultad (una por prueba)
Momento 1. Elaboración de elementos genéricos del descriptor		
Comités 1 (uno para cada prueba). Comités de Descripción de Niveles de Logro (CNL). Cada comité está compuesto por: • Coordinador de prueba • Especialistas en currículo y en investigación educativa	1.Formación 2.Grupo de discusión Análisis de los subdominios de cada área evaluada, con asignación de descriptores a cada nivel de logro, para finalmente construir una descripción global de cada nivel de logro	• Valoración de la adecuación de las etiquetas • Identificación de los descriptores de cada nivel de logro, en cada una de las áreas evaluadas • Descripción general de cada NL, incluyendo todos los subdominios del área
Fase 4. Determinación de las Puntuaciones de Corte (PC)		
Comités 2 (uno para cada prueba). Comités de Determinación de Puntos de Corte (CPC). Cada comité está compuesto por: • Coordinador de prueba • Cinco profesores(as) en ejercicio	1.Formación 2.Toma de contacto con la prueba 3.Sesiones de juicio 4.Sesión de evaluación del proceso	• Revisión de la descripción de NL realizada por el Comité 1 • Identificación de reactivos marcadores y puntuaciones de corte entre categorías o niveles de logro, para cada prueba • Valoraciones del proceso – Estudio de Validación (Hojas de Registro, Cuest. 2.1. y 2.2.)
Fase 5. Determinación de Niveles de Logro (2/2)		
Momento 2. Elaboración de las ejemplificaciones de los descriptores de los NL		
Comités 1 (uno para cada prueba). Comités de Descripción de Niveles de Logro (CNL). Cada Comité está compuesto por: • Coordinador de prueba • Especialistas en currículo y en investigación educativa	Grupo de discusión	• Ajuste de la descripción de los NL a partir de las aportaciones del Comité 2 • Redacción de ejemplos por NL • Selección de un reactivo por subdominio para cada NL, como muestra a utilizar en la fase de difusión de la información evaluativa
Comités 1	Trabajo individual	• Valoración del proceso – Estudio de Validación (Cuestionario 1)

[2] Entre ellos se encontraban quienes habrían de ejercer como coordinadores de las cuatro pruebas.

Las funciones del CNL

El Comité de elaboración de las etiquetas y descriptores de interpretación estará compuesto por un número reducido de especialistas en currículo y en investigación educativa, así como por el coordinador(a) de cada prueba, que actuará como conductor(a) del comité. El CNL debe ser independiente del comité que posteriormente trabaje en la identificación de puntuaciones de corte.

El esquema de trabajo a desarrollar del CNL se recoge en la Tabla 3.

Finalidad y objetivos del CNL

Este comité valorará la adecuación lógica de las categorías esperables de ejecución, señalando —a partir de los descriptores que componen el currículo— las características generales de la ejecución esperable en cada nivel. No se pretende que los participantes en el comité anticipen el comportamiento empírico de la muestra, sino orientar el trabajo de identificación posterior de las puntuaciones de corte desde la lógica subyacente a la construcción de la prueba. Asimismo, se trata de poder constatar si la categorización inicial que se espera realizar a partir de la prueba se basa en posibles niveles diferenciales en cuanto al contenido de la misma, de forma que no puedan darse categorías vacías o artificiales. Por último, y una vez definidas las puntuaciones de corte, se tratará de ajustar las categorías de descripción de los niveles de logro, representándolos adecuadamente y aportando muestras de ejecución de los reactivos de cada categoría o nivel de logro.

Momentos de actuación del CNL

Este trabaja en dos momentos (ver Tabla 3)

1. Previo[3] al análisis empírico de resultados; donde se desarrollan los descriptores de los niveles de logro con el fin de que sirvan de guía de contenidos para el trabajo del comité de identificación de puntuaciones de corte.

2. Posterior a la identificación de puntuaciones de corte, donde se realiza un ajuste de los descriptores, considerando los resultados obtenidos por los estudiantes de la muestra y la identificación de las puntuaciones de corte definitivas; también en ese momento se completa la descripción con ejemplos de muestra que sirvan para la posterior difusión de resultados.

Forma de trabajo del CNL

El procedimiento para su desarrollo es el de panel de discusión, en el que los miembros del comité llegan a un acuerdo acerca de las categorías de descripción de los niveles. Se trata de que en el panel se lleguen a acuerdos respecto a[4]:

[3] Como orientación se aportan las especificaciones de la prueba, así como un listado de los reactivos ordenado por dificultad. Un objetivo adicional es que se puedan detectar incongruencias en el comportamiento de las especifica-ciones, vinculadas a la especificación de los reactivos.

[4] No se trata de que los miembros del comité realicen una evaluación pormenorizada cada uno de ellos por separado de todos los descriptores del universo de medida de la prueba, emitan un juicio y se analicen las congruencias y discrepancias, buscando un sistema estadístico que sintetice la información; sino que mediante el debate, los miembros del comité establezcan acuerdos.

- La adecuación de las etiquetas propuestas para cada nivel por la Dirección de Pruebas y Medición del INEE.
- Los descriptores que pueden corresponder a cada nivel y los que puedan considerarse limítrofes o que pertenezcan a dos niveles. Para ello, se solicita a los miembros del comité que clasifiquen los descriptores en cada uno de los niveles, y se identifiquen aquellos que planteen conflicto de clasificación al no ser claramente asimilables a una sola categoría. A partir de los descriptores ya clasificados, se trata de que sinteticen el tipo de ejecución característica de cada nivel de logro. Para facilitar esta tarea se toma como referencia el análisis reticular realizado para cada materia (ver Tabla 2).
- Una vez que se dispone de la identificación de las puntuaciones de corte establecidas por el comité que mencionaremos a continuación, el Comité 1 revisa el ajuste de los descriptores utilizados, teniendo en cuenta el comportamiento empírico en la prueba, además de seleccionar los reactivos de muestra que ilustrarán la difusión de resultados.

Tabla 3. Comité 1, Momento 1. Esquema de trabajo de los comités (formato de trabajo: panel de discusión) Protocolos de actuación para el Comité 1 (CNL)

Elementos de trabajo							
• Descripción genérica de cada nivel • Propuesta de clasificación de especificaciones de subdominios, ordenados por dificultad (tablas de especificaciones) • Referencia: retícula							
Esquema de trabajo							
Momento 1.1	Trabajo cojunto del comité	• Primera valoración de la propuesta de etiquetas planteada por la dirección de pruebas del INEE • Explicación global del trabajo de la sesión.					
Momento 1.2	Trabajo por díadas			Identificación de especificaciones para el nivel Por debajo del básico	Identificación de especificaciones para el nivel Básico	Identificación de especificaciones para el nivel Medio	Identificación de especificaciones para el nivel Avanzado
		Díadas A	Subdominio 1				
		Díadas B	Subdominio 2				
		(…)	(…)				
		Díadas M	Subdominio N				
Momento 1.3	Trabajo conjunto del comité	• Segunda valoración de la propuesta de etiquetas plantada por la dirección de pruebas del INEE • Valoración del trabajo realizado para cada subdominio • Acuerdos		Descripción global del nivel Por debajo del básico, considerando todos los subdominios	Descripción global del nivel Básico, considerando todos los subdominios	Descripción global del nivel Medio, considerando todos los subdominios	Descripción global del nivel Avanzado, considerando todos los subdominios

Protocolos de actuación para el Comité 1 (CNL)

La organización del trabajo del Comité se especifica mediante protocolos de actuación —o guías de trabajo— en los que se establecen las líneas generales para conducir el desarrollo del comité. La finalidad es establecer un mismo sistema de trabajo para los comités de todas las pruebas. Los protocolos de actuación para el Comité 1 (CNL) son los siguientes:

- Para la formación,
- para la elaboración de descriptores

El tiempo de trabajo que se destina a los comités de tipo 1 es de dos días. Entre ambos momentos de trabajo se desarrolla el trabajo de los comités de tipo 2, cuya tarea es la identificación de las puntuaciones de corte en las pruebas (ver Tabla 4).

Tabla 4.
Estructura de organización de sesiones de trabajo de los Comités 1 y 2

Día	1	2	3	4	5
Mañana	Comité 1	Comité 2			Comité 1
Tarde		Comité 2		•	

Protocolo para la formación

La formación que se ofrezca al Comité 1 (CNL) para realizar su tarea debe contemplar los siguientes contenidos:

1. Finalidad y objetivos de la determinación de niveles de logro.
2. Mostrar las definiciones realizadas por otras instituciones. En cada caso, se requiere mostrar ejemplos de la materia sobre la que debía trabajar cada comité.
3. Presentar la propuesta de etiquetas del INEE, el procedimiento por el que fueron determinado, así como la definición genérica que corresponde a cada una de ellas.
4. Presentación de la forma de trabajo del Comité:
   a. Momentos 1 y 2 de actuación.
      - Momento 1: Elaboración de descriptores globales de cada nivel (previo a Comité 2).
      - Momento 2: Especificación de las características de los estudiantes en términos de competencias, habilidades y/o conocimientos, y selección de reactivos de muestra (posterior a Comité 2).
   b. Relación con el trabajo que desarrolla el Comité 2.
   c. Forma de actuación del Comité 1: Trabajará directamente con las especificaciones que se han utilizado para desarrollar los reactivos; para ello:
      - El INEE clasifica los descriptores en subdominios previamente ordenados por dificultad; esa clasificación se aportará como punto de partida para el trabajo del comité.
      - Se utilizará la retícula como referencia para identificar la posición en el currículo de cada uno de los descriptores.
      - Se presentará al conjunto del comité por parte del coordinador y se conformarán díadas, las cuales deben trabajar los subdominios completos de cada prueba, identificando los descriptores correspondientes a cada nivel.

- Desarrollados todos los descriptores por las díadas, se debatirá por el conjunto del comité cada solución. Una vez que se haya llegado a un acuerdo para cada subdominio, se redactará un párrafo que recoja las características globales de todos ellos para cada nivel. Respecto al nivel Por debajo del básico, puede definirse por exclusión, es decir, por no llegar a satisfacer las características del Básico, o bien por corresponder a competencias propias de niveles anteriores al trabajado.

5. Síntesis de la tarea a realizar.

Protocolo para la elaboración de descriptores

Momento 1 de actuación del Comité 1 (CNL). Elaboración de los elementos genéricos del descriptor

Se le presentan al Comité 1 los siguientes elementos:

a. Descripción genérica de cada nivel de competencia.

b. Propuesta de clasificación de especificaciones realizada por el INEE.

c. Se forman las díadas y se asignan los subdominios de trabajo a cada una de ellas. En los casos en que fuera posible, y con el fin de que no actúen siempre las mismas parejas en cada ocasión, las díadas se irán variando hasta trabajar el conjunto de subdominios (por ejemplo, 1-2, 3-4; 1-3, 2-4).

d. Cada díada debate sobre un subdominio y su asignación a cada nivel. Para ello, se procederá del siguiente modo:

- Se analizará la clasificación desde el nivel Básico hasta el Avanzado.
- Concluida la revisión del Básico, se analizará la clasificación del nivel Medio, enfatizando a los miembros del comité que se aseguren de que las especificaciones de ese nivel indiquen claramente un nivel diferencial respecto al Básico. Del mismo modo se procede a analizar el nivel Avanzado en relación al Medio.
- Como orientación acerca de los niveles de rendimiento que pueden corresponder a cada nivel de logro, se asumirá como guía la clasificación de verbos por niveles[5] que se recoge en la Tabla 5.

Tabla 5. Verbos asignados a niveles de rendimiento

Clasificación para niveles de rendimiento Graduación de niveles			
(1)	**(2)**	**(3)**	**Mayor(4)**
reconocer	comprender	utilizar	aplicar
encontrar	agrupar	anticipar	argumentar
identificar	asociar	predecir	criticar
nombrar	organizar	parafrasear	cuestionar
señalar	clasificar	reconstruir	opinar
elegir	jerarquizar	interpretar	reflexionar
	interpretar	resumir	valorar
		explicar	convertir
		integrar	demostrar
		solucionar	extrapolar
		cambiar	planear
			transformar

[5] Propuesta por Margarita Peon y Patricia Montero (especialistas del INEE), a partir de un estudio interno con especialistas en lenguaje.

e. Una vez alcanzado un acuerdo sobre la clasificación de especificaciones, se requiere elaborar de un párrafo que sintetice de manera global el tipo de competencias, habilidades y/o conocimientos que caractericen a los estudiantes de cada nivel.

f. Se asume que existe acuerdo cuando haya unanimidad (o asentimiento por parte de todos los miembros del comité). En ningún caso se procede a votar, pero el coordinador del comité deberá atender que no se asuman acuerdos existiendo participantes que estén claramente en desacuerdo, de forma que se mantenga el debate mientras se den posiciones diferentes, dirigiéndolo permanentemente hasta conseguir el consenso.

g. Para proceder al ajuste final de descriptores deben atenderse las recomendaciones definidas por el coordinador del seminario:

- El lenguaje debe ser técnicamente correcto y preciso, representando de forma adecuada el tipo de rendimiento característico de cada nivel.
- Debe ser comprensible para la mayor parte de personas (no sólo técnicos y especialistas, sino el profesorado, padres de familia, etcétera), de forma que facilite la comunicación de resultados a la sociedad.
- No se hará referencia a contenidos, sino a competencias, habilidades, destrezas, adquisiciones, conocimientos, maestrías, dominios, logro.
- Siempre que sea posible, se utilizarán términos que identifiquen niveles diferenciales de rendimiento en cada una de las competencias que se mencionen, siguiendo la clasificación de verbos descrita anteriormente.
- Si se incluyen en un mismo nivel diversos tipos de competencias o habilidades, se hará referencia explícita a cada una de ellas, identificando sus niveles de rendimiento.
- En el caso en que se incluyan términos o palabras que puedan ser interpretables, se añadirán sinónimos o los elementos necesarios que sirvan para aclarar su significado exacto.
- No se utilizarán en ningún caso palabras o términos susceptibles de interpretaciones peyorativas o discriminatorias.
- Se utilizará un lenguaje no sexista y respetuoso con la diversidad de razas, credos y circunstancias personales y sociales.

Momento 2 de actuación del Comité 1 (CNL). Elaboración de las ejemplificaciones del descriptor

Una vez que se disponga de las puntuaciones de corte identificadas por el Comité 2, el Comité 1 recibirá la información acerca de los reactivos que correspondan a cada nivel, así como de las especificaciones de los mismos, para que, a partir de esta información:

a. Se compare la definición original que aprobó el Comité 1 con la resultante del Comité 2, de forma que puedan identificarse las discrepancias entre ambas propuestas.

b. Identificadas las diferencias en cada nivel, se valore el grado en que afectasen la redacción del descriptor, para ajustar de forma precisa a la clasificación que emane de la identificación de puntuaciones de corte realizada por el Comité 2.

c. Posteriormente, el Comité 1 redactará las ejemplificaciones (ver Tabla 4) correspondientes a cada nivel de logro. Para este cometido, se incluirá una frase o descripción que ejemplifique las características de los estudiantes del nivel correspondiente en cada una de las competencias del mismo. Para ello, deben tomarse como referencia las especificaciones de los reactivos y los reactivos mismos, agrupándolos en su descripción en relación al tipo de competencia o habilidades a que correspondían.

d. Finalmente, se seleccionará un reactivo de cada subdominio como muestra del tipo de tareas que los estudiantes pertenecientes a cada nivel pueden realizar. Esta selección

debe ser muy cuidadosa, pues se trata de identificar los reactivos que se publicarán como ejemplificaciones en los informes y comunicados institucionales. Para elegirlo, se tendrá en cuenta:

- La dificultad del reactivo para el grupo de estudiantes del nivel, que no debe ser inferior al 67%, aspecto que, de hecho, asegura el procedimiento seguido en la determinación de la puntuación de corte.
- Su representatividad respecto a las competencias que caracterizan al nivel.

e. Como en el caso de la redacción de descriptores, hay que tener en cuenta las consideraciones realizadas anteriormente en relación al lenguaje a utilizar.

2.3 Determinación de las puntuaciones de corte

El procedimiento para determinar las puntuaciones de corte que sirvan para identificar los niveles de logro se basa en una adaptación del método Bookmark, considerando especificaciones respecto al formato de juicio derivadas de variaciones del método de Angoff, y consideraciones empíricas de los métodos utilizados por De la Orden (1998) y por Gaviria y Tourón (2000).[6] Así, se sustenta sobre la actuación de un comité de expertos que juzgan, a partir de los resultados obtenidos en los reactivos del Excale, cuáles son los elementos de la prueba característicos de cada nivel de logro. La especificación del procedimiento pasamos a comentarla a continuación.

Comité de determinación de Puntuaciones de Corte (CPC)

El comité de determinación d e niveles de logro estará compuesto por profesores en ejercicio, conocedores del funcionamiento real de la materia y de los alumnos tipo. El número de miembros del comité será de cinco participantes. En este comité, como en el caso anterior, el director de prueba fungirá como coordinador del mismo.

Formato de trabajo del comité: proceso

El trabajo del comité se organiza de la siguiente manera:

1. Sesión de formación: Desarrollada por el conductor del seminario —en sesión plenaria de los cuatro comités— y por el coordinador de prueba —en sesión interna de cada comité—; su finalidad es explicar a los miembros del comité los objetivos de su trabajo, así como los procedimientos a seguir para la emisión de juicios.

2. Toma de contacto con la prueba: Con el fin de familiarizarlos con el Excale, los miembros del comité responderán una prueba completa; con ello se pretende que comprueben la calidad, claridad, niveles de dificultad, etc., de los reactivos que posteriormente juzgarán.

3. Sesiones de juicio: Se establecerán tres sesiones de juicio, en las que se aporte retroalimentación al comité acerca de sus niveles de acuer-do/congruencia, así como en relación a las consecuencias de la aplicación de los niveles identificados. Los objetivos de esta estrategia son: facilitar la congruencia final en torno a los niveles identificados; identificar expertos que ofrezcan valoraciones "extremas", y ajustar de forma realista los niveles resultantes.

4. Información de retroalimentación a participantes: Un problema a tener en cuenta es que se trata de identificar varias puntuaciones de corte a lo largo de un continuo. De esta forma, se prevén dos tipos de discrepancias:

[6] Las tareas complejas abiertas, como por ejemplo, composiciones escritas podrían valorarse de forma independiente a través de un método holista. Ver referencias en bibliografía del Anexo 1.

- Entre jueces; se trata de identificar la congruencia o discrepancia en los juicios, de forma que ello actúe como elemento de reflexión a los expertos que se aparten significativamente del conjunto de estimaciones. Esta estrategia es frecuente y persigue estimaciones más robustas y representativas.
- En las puntuaciones de corte pueden darse las siguientes discrepancias entre las valoraciones emitidas: a) diferencias generalizadas, cuando no se dan acuerdos en la identificación de ninguno de los niveles, y b) localizadas en uno o algunos niveles. En cada caso, el coordinador del comité debe aportar información dirigida a solventar los problemas encontrados.

En cualquier caso, la información de retroalimentación a participantes se dirigirá a los siguientes aspectos:

- Grado de congruencia entre los expertos para cada puntuación de corte, mediante indicadores univariados para cada puntuación de corte en cada sesión de juicio
- Número de reactivos que definan cada nivel de logro, discrepancias en la identificación de reactivos entre jueces
- Distribución porcentual de sujetos en cada nivel de logro

5. Sesión de evaluación del proceso: Una vez concluido el proceso, se recogerán informaciones de los participantes acerca del desarrollo del mismo.

Formato de juicio

Para la emisión de juicio se requiere que cada miembro del comité trabaje sobre un cuadernillo de reactivos ordenados (CIO), el cual incluya los reactivos del Excale ordenados de menor a mayor dificultad. Cada reactivo se presentará completo e identificando su nivel de dificultad.

La tarea a plantear al comité será identificar qué reactivos pertenecen a cada una de las categorías, comenzando por el reactivo más fácil y por la categoría de menor nivel de logro. Para ello, deben los participantes examinar cada uno y responder si un sujeto de la categoría que se esté valorando en ese momento será capaz de responder correctamente al mismo.

Así, la pregunta a responder para cada reactivo es "¿un sujeto del nivel θ puede responder correctamente este reactivo?". La respuesta que debe dar cada participante es en términos de SÍ/NO. No obstante, deberá especificarse a los participantes que la pregunta no se refiere a si todos los sujetos son capaces de hacerlo, sino a si la mayoría de los sujetos de dicho nivel lo serán, tomando como referencia una probabilidad de al menos el 67% de ellos.

El cambio de nivel de logro se producirá cuando surge un reactivo del que se entiende es razonable que un sujeto del nivel actual no lo pueda responder. Ese reactivo actúa entonces como marcador para identificar la puntuación de corte. En caso de que se identifique un reactivo de cambio de nivel, pero posteriormente exista duda acerca de si algunos reactivos de mayor dificultad podrían ser correctamente resueltos por sujetos del nivel anterior (hecho poco posible, pero probable), cada participante debe revisar su marcador hasta asegurarse que esté situado en el reactivo más representativo, según su opinión. No obstante, cada participante debe identificar tantos reactivos marcadores como puntuaciones de corte entre niveles. Como la categorización es de cuatro niveles, se tendrán que identificar tres marcadores (ver Figura 1 y Tabla 6).

Figura 1.
Ilustración de cuadernillo de reactivos ordenados y reactivos marcadores

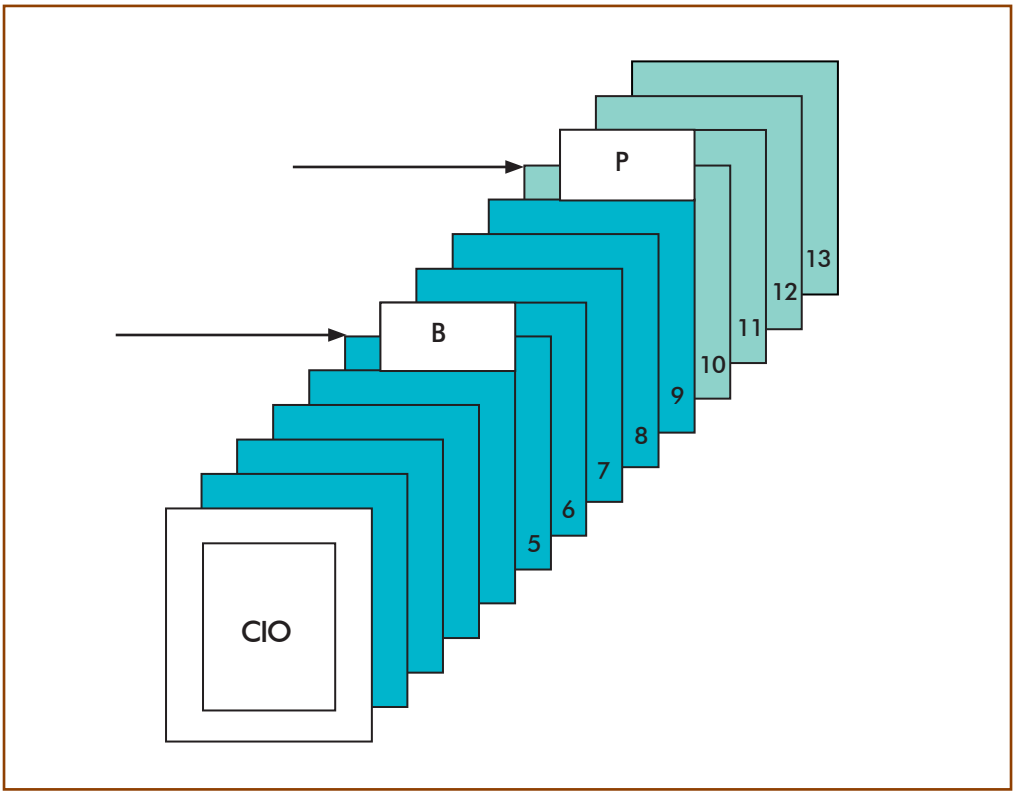
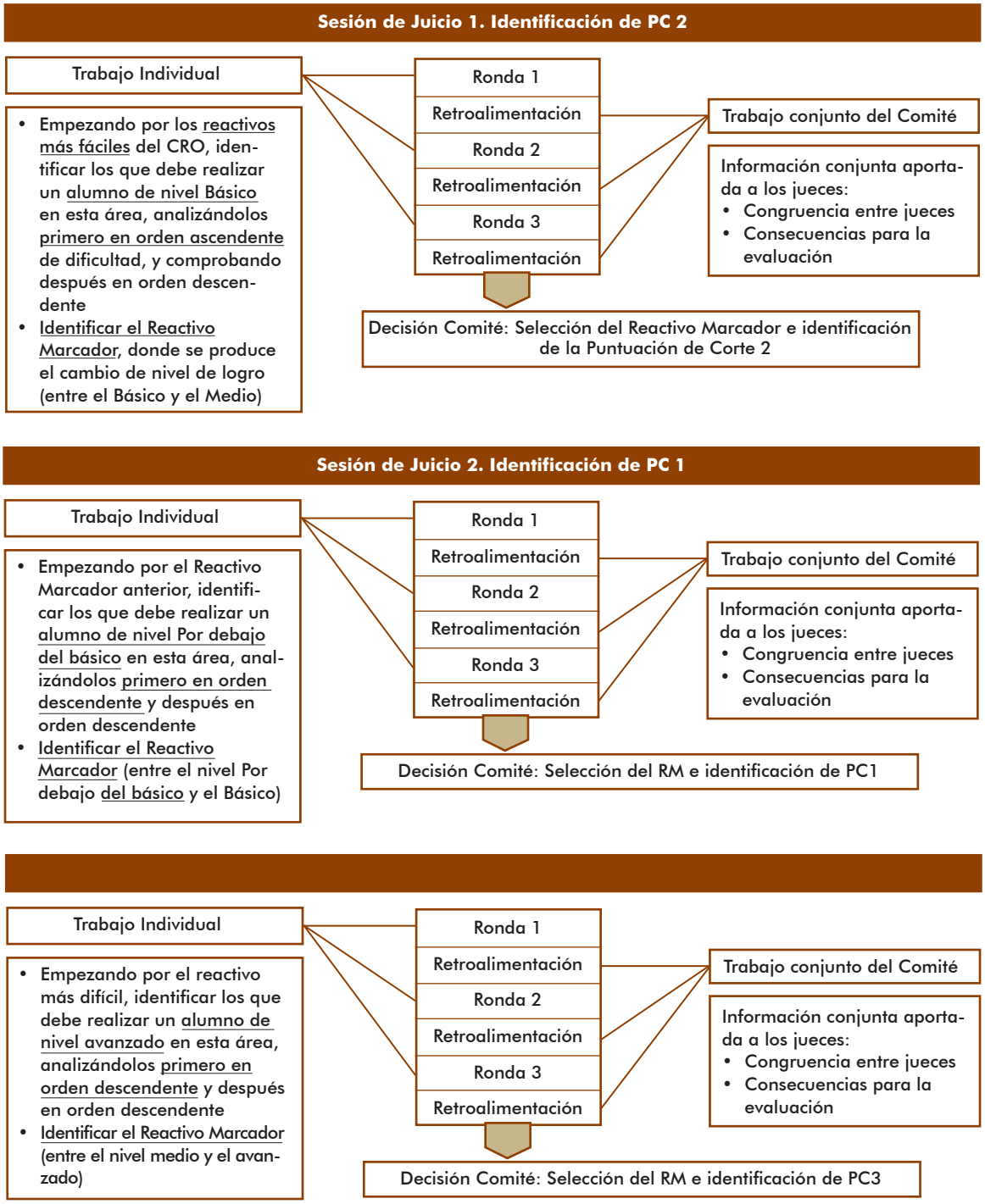


Tabla 6.
Esquematización de niveles y puntuaciones de corte

Niveles	Puntuaciones de corte
I. Por debajo del Básico	
II. Básico	θ_1
III. Medio	θ_2
IV. Avanzado	θ_3

Una esquematización del proceso de juicio, en la que se indican tanto las tareas a realizar por los participantes como el sistema de rondas de juicio, se presenta en la Tabla 7.

Tabla 7. Comités 2 (CPC). Sesiones de Juicio



Criterios de decisión e identificación de puntuaciones de corte

Teniendo en cuenta los desarrollos de los que se informa en la literatura especializada, con tres sesiones de juicio normalmente se llega a niveles representativos con un elevado nivel de congruencia entre los participantes. No obstante, si ello no ocurriera, sería necesario revisar los elementos del procedimiento que provocan la falta de acuerdo, con el fin de tomar decisiones sobre el proceso y respecto a los niveles resultantes. Por este motivo, se tomará como esquema de trabajo realizar tres rondas de juicio (ver Tabla 7).

Las puntuaciones de corte están representadas por la mediana del nivel de habilidad correspondiente a los reactivos identificados como marcadores. El objetivo es identificar como puntuación de corte aquella la resultante del máximo nivel de congruencia en la opinión y, por lo tanto, se pueda entender como un estimador robusto del consenso intersubjetivo.

Como criterios de calidad para considerar cerrado el proceso se tendrán en cuenta los siguientes:

- Los niveles de congruencia entre los participantes para cada puntuación de corte.
- La valoración por parte de los participantes acerca de la representatividad de los niveles obtenidos en cuanto al porcentaje de sujetos identificados en cada nivel.

Así, se pretende que las puntuaciones de corte estén sustentadas en un alto nivel de congruencia interjueces, y sean representativas de la realidad escolar (según la opinión del comité).

Protocolos de actuación para el Comité 2 (CPC)

Protocolo para la formación

La formación que se dé al Comité 2 (CPC) para realizar su tarea, debe contemplar los siguientes contenidos:

1. Explicación de la finalidad y objetivos de la determinación de niveles de logro.
2. Mostrar las definiciones realizadas por otras instituciones; en cada caso, con ejemplos de la materia sobre la que trabajaba cada comité.
3. Presentación de la propuesta de niveles y descriptores que realiza el INEE, explicando el proceso que se ha desarrollado hasta el momento y las labores realizadas por la institución a través de sus técnicos y comités.
4. Presentación del formato de juicio, indicando claramente el procedimiento a seguir, especificando que se realizarían diversas rondas de juicio para la determinación de cada puntuación de corte, el tipo de retroalimentación que se ofrecerá y los criterios de convergencia y control se van a utilizar.
5. Muestra de un ejemplo simulado.
6. Presentación del CRO (Cuadernillo de Reactivos Ordenados)[7] y explicación de la información que corresponda a un reactivo.

[7] Este cuadernillo incluye los reactivos del Excale ordenados de menor a mayor dificultad. Cada reactivo se presenta completo e identificando su nivel de dificultad.

7. Concluida la sesión de formación, se les administrará la prueba sobre la que se van a identificar las puntuaciones de corte. La finalidad es que tengan contacto real con la prueba. A los miembros del comité se les presentará una prueba completa, para que la cumplimenten. Posteriormente se les aportará la clave de respuestas con el fin de comprobar si fallaron en algún reactivo, y así tener la oportunidad de comprobar la calidad, claridad y niveles de dificultad de los reactivos que juzgarán posteriormente.

Protocolos para las sesiones de juicio

1. A cada miembro del comité se le facilitará un cuadernillo de reactivos ordenados (CRO).
2. Cada miembro emitirá su juicio individualmente a partir de la revisión del CRO.
3. Los jueces deben tener como referencia el descriptor del nivel y, en su caso, podrá aceptarse un debate acerca de sus implicaciones.
4. Cada sesión de juicio estará centrada en identificación de una única puntuación de corte, con la secuencia que posteriormente describiremos. Para entonces deben haberse realizado hasta tres rondas de juicio, con el fin de facilitar la convergencia entre los jueces.[8] Al finalizar cada ronda, se introducirán los datos y se realizará un breve análisis para comprobar:
   a. la convergencia entre jueces
   b. las consecuencias de la aplicación de la puntuación de corte para describir los resultados de la evaluación

   Dicha información se ofrece como feedback (cuyo protocolo también describiremos posteriormente) a los participantes que, a continuación, podrán revisar su juicio anterior.
5. La tarea que se plantea a los participantes es identificar, comenzando por el reactivo más fácil y por la categoría de menor nivel de logro, qué reactivos pertenecen a cada una de las categorías. Para ello, deberán examinar cada reactivo y responder si un sujeto de la categoría que se estaba valorando en ese momento sería capaz de responder correctamente al reactivo (tal como lo describimos en Anexo 1).
6. La pregunta que se debe responder para cada reactivo es: "¿un sujeto del nivel θ_1 puede responder correctamente este reactivo?".[9] La respuesta que debe dar cada participante es SÍ/NO, como hemos indicado anteriormente.
7. En términos generales, la identificación del reactivo en el que se produzca el primer "NO" indica el cambio de nivel de logro: es justamente el reactivo marcador, que se produce cuando el juez identifica un reactivo que estima es poco probable lo pueda responder un sujeto del nivel actual. En otras palabras, ese reactivo actúa como marcador para identificar la puntuación de corte, la cual representará el nivel mínimo que deben mostrar los alumnos para ser considerados dentro de un nivel de logro determinado.
8. En caso de que se identifique un reactivo como marcador, pero posteriormente se dude si algunos reactivos de mayor dificultad podrían ser correctamente resueltos por sujetos del

[8] Salvo que se produzca la convergencia en un menor número de rondas, o que se advierta la conveniencia de desarrollar alguna ronda más.

[9] No obstante, se debe tener cuidado, pues su aplicación puede causar equívocos. Por ejemplo, la θ_1 que separa los niveles Básico y Por debajo del básico era conveniente que se base en la identificación de todos los reactivos que debe conocer un alumno para considerar que tiene un nivel Básico. Obviamente, por debajo de la puntuación correspondiente se sitúa el nivel inferior.

nivel anterior —o viceversa— (hecho poco posible, pero probable), cada participante revisará su marcador hasta asegurarse que está situado en el reactivo más representativo, según su opinión. De los reactivos que se identifiquen con estas características, es necesario que:

a. queden señalados por cada juez.
b. se analice el posible motivo de su mala ubicación.
c. sean recogidos como "incidencias en la determinación de estándares".

Estos casos pueden dar origen a recomendaciones como consecuencia de la evaluación, pues podrían constituir interpretaciones complementarias de carácter cualitativo a la información meramente cuantitativa.

9. Este tipo de incidencias pueden darse por diversos motivos, entre ellos podemos señalar los más frecuentes:

a. un contenido teóricamente fácil medido por un reactivo mal diseñado, de forma que ello lo convierta en más difícil.
b. un contenido que no se imparta habitualmente en las clases, aunque esté presente en el currículo (por ejemplo, los contenidos de Estadística en cursos de Matemáticas).
c. un contenido teóricamente más difícil puede aparecer como más fácil cuando el reactivo que lo mide tiene mal diseñados los distractores o incluye pistas que orientan hacia la identificación de la respuesta correcta. Este tipo de incidencias se deben analizar convenientemente con el fin de extraer las consecuencias oportunas.

10. Dado que se trata de identificar tres puntuaciones de corte, y considerando que cada nivel de logro debe responder a competencias claramente delimitadas, la secuencia de identificación de puntuaciones se altera (desde la mínima a la máxima), siguiendo el siguiente esquema de trabajo:

θ_2, θ_1 y θ_3

donde: θ_2 es la puntuación de corte entre el nivel Básico y el Medio

θ_1, es la puntuación de corte entre el nivel Por debajo del básico y el Básico

θ_3, es la puntuación de corte entre los niveles Medio y Avanzado

Esta secuencia facilita la tarea a nivel cognitivo y conlleva a que cada juez explore la identificación del reactivo marcador en dos direcciones (primero ascendente y luego descendente).

La tarea para cada puntaje es:

a. θ_2 se identifica desde los reactivos más fáciles, en orden ascendente (posteriormente se comprueba en sentido inverso). La tarea que deben realizar los participantes es identificar todos los reactivos que debe realizar un alumno cuyo nivel es Básico.

b. θ_1 corresponde al nivel mínimo de competencia del nivel Básico. Se identifica en sentido descendente, partiendo del reactivo marcador señalado con anterioridad (posteriormente se comprueba en sentido inverso). La tarea es identificar el nivel mínimo que debería exigirse para poder valorar a un alumno como perteneciente al nivel Básico.

c. θ_3 es la puntuación de corte que separa los niveles Medio y Avanzado. Se procede en orden descendente; es decir, los reactivos más difíciles a los más fáciles (posteriormente se comprueba en sentido inverso). La tarea que se planteará a los participantes será determinar si este reactivo pueden responderlo únicamente los alumnos de nivel Avanzado.

11. Una vez obtenida la convergencia en torno a una puntuación de corte se dará por concluida la sesión, dando paso a la sesión correspondiente a la siguiente puntuación, hasta concluir el proceso.

12. Cada participante debe identificar tantos reactivos marcadores como puntuaciones de corte entre niveles. De modo que, como la categorización es de cuatro niveles, deben identificase tres puntuaciones de corte, es decir, tres marcadores, tal como indicamos anteriormente y se muestra en el esquema de la Tabla 12.

Protocolo para la retroalimentación

1. La información de retroalimentación tiene por objeto ayudar a orientar el acuerdo inter-jueces. No se trata de dirigir el juicio, sino de aportar elementos de reflexión para que cada uno, de manera individual, pueda revisar su opinión y, si lo estima conveniente, modificarla.

2. El formato con que se produjo la información reúne los siguientes elementos (ver Cuadro 10).

   a. Congruencia entre jueces. Se indica:

      - El rango de puntajes que se hayan emitido (por ejemplo, entre 12 y 15).
      - La variabilidad de los juicios expresada a partir de la desviación típica.
      - El nivel de acuerdo que haya entre los jueces en ese momento, mostrado a partir de su distribución gráfica.
      - La distancia entre las puntuaciones de corte señaladas por cada juez.

   b. Consecuencias para la evaluación. Se indica:

      - El porcentaje de sujetos que quedarían por encima y por debajo de cada puntaje y de la posible puntuación de corte, estimada como la media de las aportadas.
      - Distancia en puntajes y en porcentaje de sujetos desde la puntuación de corte y la media.

   c. En la sesión conjunta no se facilita información acerca de los juicios emitidos por cada uno de los jueces. En todo caso, se comenta con cada juez, de forma privada si así lo demandan, aspectos relativos a sus juicios.

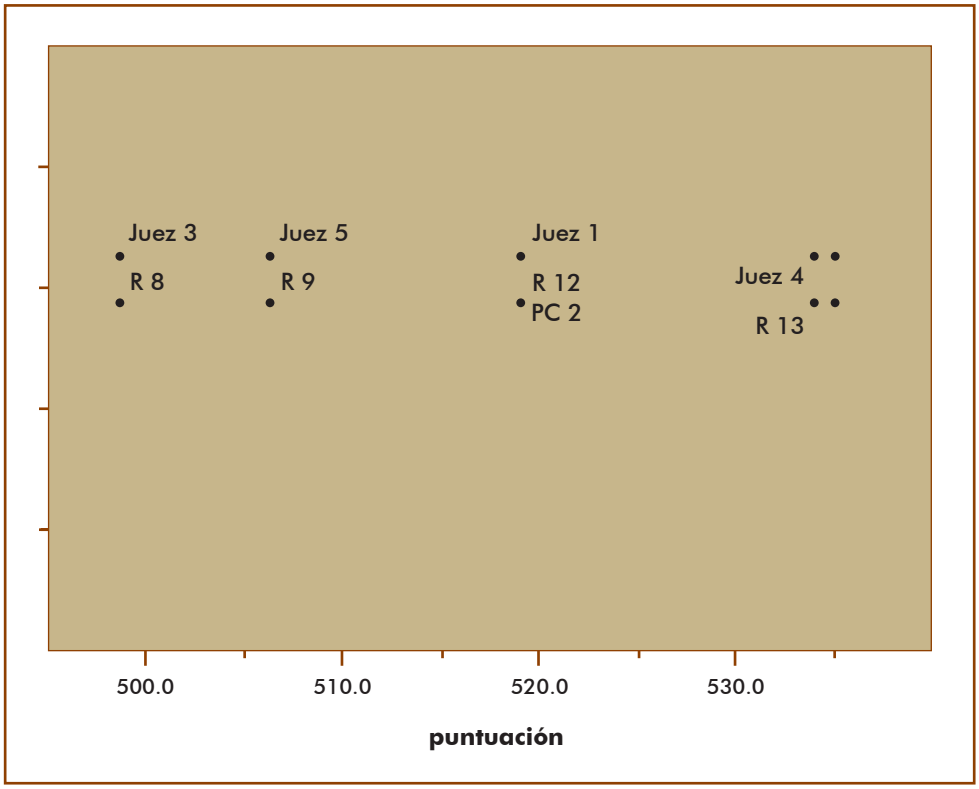
En la Tabla 8 y en las gráficas 1 y 2 se ejemplifica el tipo de información de retroalimentación aportada.

Tabla 8. Muestra de informaciones para la retroalimentación del Comité 2

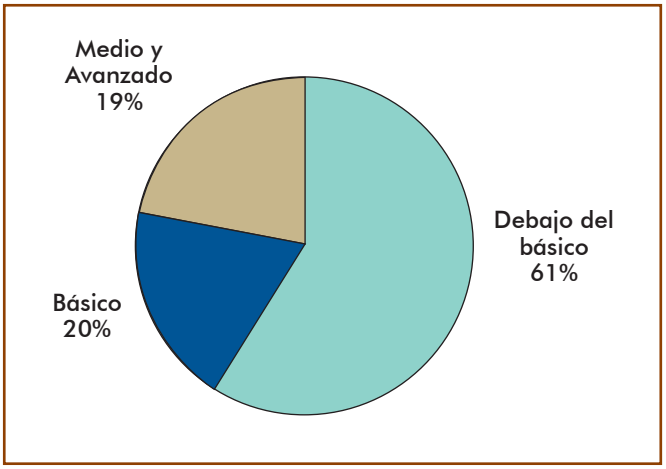
Asignatura		Nivel	
Asistencia	**Juez**	**Reactivo marcador**	**Puntuación de corte**
S	Juez 1	12	518.9
S	Juez 2	14	535
S	Juez 3	8	498.7
S	Juez 4	13	533.7
S	Juez 5	9	506.3

		Ronda: 3
Indicador	**Reactivo marcador**	**Puntuación de corte**
PC1	12	518.9
Mínimo	8	498.7
Máximo	14	535
Desv. Están.		16.2

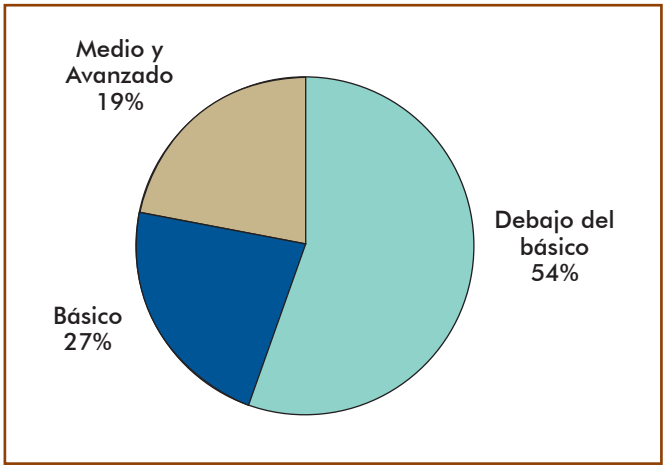
Gráfica 1. Distribución de los puntos de corte PC1 sugeridos



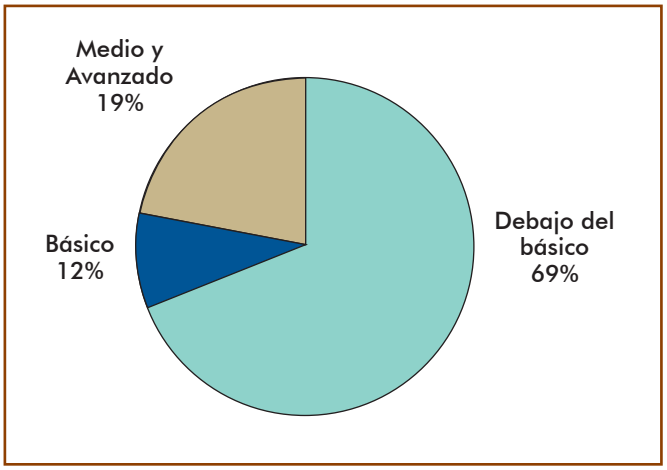
Gráfica 2. Distribución de los puntos de corte PC1 sugeridos



Con PC1 mínimo



Con PC1 máximo



Protocolo para la selección de las puntuaciones de corte

1. Las puntuaciones de corte corresponden a la mediana del nivel de habilidad correspondiente a los reactivos identificados como marcadores.

2. Como criterios de calidad para considerar cerrado el proceso se tendrán en cuenta:

   a. Los niveles de congruencia entre los participantes para cada puntuación de corte. Para este cometido, se toma como referencia la distancia entre las puntuaciones de corte y la desviación de los juicios emitidos.

   b. La valoración por parte de los participantes acerca de la representatividad de los niveles obtenidos en cuanto al porcentaje de sujetos identificados en cada nivel.

   c. En cualquier caso, el ajuste final de cada puntuación se realizará de forma que se mantenga una distancia suficiente entre puntuaciones de corte. Este aspecto, así como otros indicadores de calidad, se muestran en el apartado de validación.

3 VALIDACIÓN DE LOS NIVELES DE LOGRO

La validación de los estándares de interpretación —o niveles de logro— se basa en diversas perspectivas, sustentándose sobre evidencias extraídas tanto acerca del proceso de determinación de estándares utilizado, como acerca de los niveles de logro considerados en sí mismos. En este sentido, la validación de los niveles de logro del Excale también se ha abordado desde estas consideraciones. Comentaremos a continuación las características de las dos perspectivas de validación que se han tenido en cuenta y que se incluyen en este informe.

3.1 Validación a través de la evaluación del proceso de determinación de los niveles de logro

Buena parte de los problemas de algunos estándares es la falta de credibilidad acerca del proceso por el que se han desarrollado. En este sentido, la evaluación del proceso de determinación de los niveles de logro actúa como una primera evidencia de validez de los mismos, asegurando la transparencia del proceso y su replicabilidad. Es por ello que la documentación exhaustiva de todo el proceso debe atenderse como primera garantía. Para la descripción del plan de evaluación que se desarrolló para la validación del proceso, se revisaron los diversos componentes del mismo.

Objeto y finalidad de la evaluación

Se trata de evaluar si el proceso de determinación de los niveles de logro se ha desarrollado de forma adecuada. La finalidad es doble:

- formativa, de manera que durante el proceso se intenten corregir problemas detectados durante el desarrollo del mismo.
- sumativa, como rendición de cuentas (y evidencia de validez) acerca de la representatividad y calidad de los niveles de logro identificados como sistema de interpretación de puntuaciones.

Realización de la evaluación, informes, control y participantes

Para cumplir con ambas finalidades, el proceso de evaluación contempla dos etapas, como se muestra en la Tabla 9.

1. El Informe de Evaluación será dirigido y desarrollado por un equipo externo de evaluación, encargado de recoger la información, analizarla y sugerir junto al coordinador de prueba las mejoras a lo largo del proceso, así como elaborar un informe final de la evaluación del mismo.

   Este equipo participará en las sesiones a través de un observador externo que funja como asistente del coordinador de prueba, si bien su rol debe centrarse exclusivamente en las tareas de evaluación.

2. Comité Meta-Evaluador[10] (CME), encargado de revisar la información derivada de la evaluación, así como la documentación técnica del proceso. Su tarea es comprobar, a través del informe, la adecuación general del procedimiento, tanto en la determinación de niveles de logro como en la evaluación. Como resultado de su actuación, emitirá un informe final de validación del proceso, en el que aportará las evidencias necesarias como apoyo o refutación de la calidad del procedimiento desarrollado.

En cualquier caso, todos los miembros del comité, así como el coordinador de prueba, son también participantes en la evaluación, aportando información y valoraciones a través de todo el proceso.

Tabla 9. Síntesis de informes a emitir y unidades de trabajo encargadas

Informes a emitir	Quién desarrolla la tarea		
	Equipo extetrno	Comité CNL	Comité CME
Informe de evaluación del proceso	Dirige, recoge y elabora información	Debate del informe, emisión de sugerencias	Revisa y analiza: • Documentación técnica • Informe • Metodología seguida en el proceso
Informe de validación del proceso			Valida o refuta el procedimiento

Plan de trabajo: fuentes de información, variables e indicadores, momentos de recopilación de información, instrumentos y análisis

Las fuentes de información a considerar son:

a. Los participantes, en al menos tres aspectos:
   - El análisis de sus respuestas de identificación de puntuaciones de corte.
   - El conocimiento y comprensión de los métodos y procedimientos a utilizar.
   - Sus opiniones acerca del proceso.

b. El coordinador de prueba: sus valoraciones acerca del proceso.

c. El observador externo: sus valoraciones acerca del proceso.

d. El Comité Meta-Evaluador (CME): sus valoraciones metodológicas (validación del informe de evaluación) deberán incorporarse en un informe final del INEE, junto a este informe.

Las variables e indicadores a tener en cuenta para este proceso, así como las fuentes de información de las que se extrajeron aparecen en la Tabla 10.

Respecto a los momentos de recopilación de información, las variables e indicadores de entrada se recabaron previamente o al inicio de las sesiones de juicio. Hubo que distinguir entre indicadores y variables de proceso, así, el indicador relativo a la comprensión de la tarea y procedimientos se recogió al finalizar la sesión de formación, previamente a iniciarse las sesiones de juicio; los indicadores de cambio y de satisfacción son subsiguientes a las sesiones de juicio, en este caso se extrajeron tres medidas.

[10] Podría asumir este rol el grupo asesor, o bien el Consejo Técnico.

Por último, en relación a los indicadores contextuales, hay que señalar que la comparación del funcionamiento de los diversos comités de las diferentes materias se incluye como resultado del análisis realizado en este informe.

Respecto a los instrumentos utilizados se pueden realizar las siguientes consideraciones:

a. Variables de entrada, recogidas mediante cuestionario dirigido a los participantes en los comités.
b. Valoración del conocimiento y comprensión de tareas y métodos, mediante prueba estandarizada dirigida a los miembros de los comités.
c. Tasas de cambio en indicadores y niveles de congruencia, fiabilidad, características de las distribuciones, tanto las variables de proceso como las de producto: datos de carácter estadístico, tanto a nivel univariado como multivariado.
d. Valoración del funcionamiento del comité en sus diferentes facetas: cuestionario dirigido a los miembros del comité y registro observacional dirigido a coordinador de prueba y observador externo.

Por último, una vez completado el informe, el CME comprobará la adecuación de la información y las valoraciones contenidas en él, de forma que deberá emitir un juicio valorativo global que atienda al menos a tres aspectos:

1. la calidad del proceso de determinación desarrollado
2. la adecuación metodológica y de las decisiones tomadas
3. la calidad de los niveles de logro identificados

Para esta evaluación, el CME dispondrá de toda la documentación disponible acerca de los trabajos realizados por el CSI y CNL, así como del informe de evaluación, y podrá recabar la información complementaria que precise, tanto documental como de audiencias con participantes.

3.2 Validación a través de la evaluación del resultado: la calidad de los niveles de logro como sistema de interpretación

Como es bien sabido, la validación de cualquier prueba se constata a través del tiempo, la investigación y uso de las pruebas, y se basa en la acumulación de evidencias que demuestren la calidad, credibilidad y utilidad de la información extraída a partir de ellas. En este espacio se aborda genéricamente el problema; en el Anexo 1 se profundiza en él.

La validez de los niveles de logro como sistema de interpretación de los resultados de las pruebas dependen de múltiples factores, y el INEE, como responsable último de la evaluación, es quien debe priorizar cuáles son las evidencias fundamentales a recoger para utilizar de manera óptima la información evaluativa.

Tabla 10. Síntesis de variables/indicadores y fuentes de información

Tipo	Variables / Indicadores	Fuentes de información				
		Valoraciones de niveles de logro	Opiniones de participantes	Coordinador de prueba	Observador externo	Otras fuentes
De entrada	Características profesionales de los participantes[11]		x	x		
De proceso	Comprensión de la tarea y de los procedimientos a utilizar	x	x	x	x	
	Número de sesiones de juicio			x	x	
	Cambios en la identificación de puntuaciones de corte de una a otra sesión de juicio	x				
	Cambios en la fiabilidad asociada a las puntuaciones de corte de una a otra sesión de juicio	x				
	Cambios en la distribución porcentual de los sujetos a partir de los niveles de logro identificados de una a otra sesión de juicio	x				
	Satisfacción con el proceso de formación	x	x	x	x	
	Satisfacción con los procedimientos utilizados	x	x	x	x	
	Satisfacción con el funcionamiento global del comité	x	x	x	x	
De resultado	Congruencia en la identificación de puntuaciones de corte (en cada sesión de juicio).Perspectivas univariada y multivariada	x				
	Fiabilidad asociada a las puntuaciones de corte en cada nivel	x				
	Distribución porcentual de los sujetos en los niveles de logro	x				
	Satisfacción con la adecuación de los niveles de logro determinados	x	x	x	x	
De contexto	Comparación del funcionamiento de los diversos comités de las diferentes materias					x
	Análisis lógico de los niveles de logro identificados para cada materia con los utilizados en otro proyectos evaluativos comparables					x

El INEE deberá determinar, entre los diversos estudios programados, cuáles son necesarios para asegurar un uso adecuado de las puntuaciones. No obstante, y como una primera aproximación a la validación del producto, se aporta un estudio dirigido a analizar la convergencia de los perfiles identificados a partir de análisis de conglomerados (de carácter empírico) con la clasificación derivada de la aplicación de los estándares.

[11] Se recabará información de síntesis acerca del historial profesional de los miembros del comité, así como en relación a su actuación evaluadora como profesor(a).

BIBLIOGRAFÍA

- 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.
- Andrew, B.J. y Hecht, J.T. (1976): A preliminary investigation of two procedures for setting examination standards. *Educational and Psychological Measurement, 36*, 35-50.
- Angoff, W.H. (1971): Scales, norms, and equivalent scores. In R.L. Torndike (Ed), *Educational measurement* (pp. 508-600). Washington, DC: American Council on Education.
- Backhoff, E. (2005): *Exámenes de Calidad y Logro Educativos (Excale): Proceso de construcción y características básicas. Los Temas de la Evaluación*, Colección de folletos. México: Instituto Nacional para la Evaluación de la Educación.
- Behuniak, P., Archambault, F.X. y Gable, R.K. (1982): Angoff and Nedelsky standard setting procedures: Implications of the validity of proficiency test score interpreting. *Educational and Psychological Measurement, 42, 1*, 247-255.
- Berk, R.A. (1986): A consumer's guide to setting performance standards on criterion referenced tests. *Review of Educational Research, 56, 1*, 137-172.
- Berk, R.A. (1996): Standard setting: the next generation (Where few psychometricians have gone before). *Applied Measurement in Education, 9 (3)*, 215-235.
- Beuck, C.H. (1984): A method for reaching a compromise between absolute and relative standards in examinations. *Journal of Educational Measurement, 21*, 147-152.
- Block, J.H. (1978): Standards and criteria: A response. *Journal of Educational Measurement, 15 4*, 291-295.
- Brown, W.J. (2001): Social, educational, and political complexities. In G.J. Cizek (Ed), *Setting performance standards: Concepts, Methods, and Perspectives* (pp. 373-386). Mahwah, NJ: Erlbaum.
- Camilli, G., Cizek, G.J. y Lugg, C.A. (2001): Psychometric theory and the validation of performance standards: History and future perspectives. In G.J. Cizek (Ed), *Setting performance standards: Concepts, Methods, and Perspectives* (pp. 445-476). Mahwah, NJ: Erlbaum.
- Carson, J.D. (2001): Legal issues in standard setting for licednsure. In G.J. Cizek (Ed), *Setting performance standards: Concepts, Methods, and Perspectives* (pp. 427-444). Mahwah, NJ: Erlbaum.
- Castro, M (2001): How accurate are writing performance assignment raters? 2001 LAUSD rater reliabity study. CSE Technical Report (CRESST, UCLA).
- Chinn, R.N. y Hertz, N.R. (2002): Alternative approaches to Standard setting for licensing and certification examinations. *Applied Measurement in Education, 15*, 1-14.
- Cizek, G.J. (1993): Reconsidering standards and criteria. *Journal of Educational Measurement, 30* (2), 93-106.
- Cizek, G.J. (1996a): Setting passing scores. *Educational Measurement: Issues and Practice, 15* (2), 20-31.
- Cizek, G.J. (1996b): Standard setting guidelines. *Educational Measurement: Issues and Practice, 15(1)*,12,13-21.

- Cizek, G.J. (2001a): More unintended consequences of high-stakes testing. *Educational Measurement: Issues and Practice, 20 (4)*, 19-27.
- Cizek, G.J. (2001b): 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.
- Clauser, B.E. y Clyman, S.G. (1994): A contrasting-groups approach to standard setting for performance assessments of clinical skills, *Academic Medicine, 69, 10*, 42-44.
- Cross, L. H., Impara, J. C., Frary, R. B. y Jaeger, R. M. (1984): A comparison of three methods for establishing minimum standards on the National Teacher Exarninations. *Journal of Educa¬tional Measurement, 21*, 113-130.
- De Gruijter, D.N. (1985): Compromise methods for establishing examination standards, *Journal of Educational Measurement, 22*, 263-269.
- De la Orden, A. (1985): Hacia una conceptualización del producto educativo. *Revista de Investigación Educativa, 3*, 6, 271-284.
- De la Orden, A. (1993) La escuela en la perspectiva del producto educativo. Reflexiones sobre la evaluación de centros docentes. *Bordón*, 45 (3), 263-270.
- De la Orden, A. (1995) *Hacia un modelo para evaluar la calidad universitaria*. Ponencia en el Seminario sobre Evaluación de la Calidad Universitaria, Centro Anáhuac de Investigación y Servicios Educativos, México
- De la Orden, A. (2000): Estándares en la evaluación educativa. Ponencia presentada en las primeras *Jornadas de Medición y Evaluación* (marzo, 2000). Valencia: Universidad de Valencia.
- De la Orden A., Bisquerra R., Gaviria J.L., Gil G., Jornet J.M., López Freire F.A., Sánchez Diaz J., Sánchez Villafaina M.C., Sierra J. y Tourón F.J. (1998): Los resultados escolares. Diagnóstico del Sistema Educativo, 1997. Madrid: Ministerio de Educación y Cultura, Secretaría General de Educación y Formación Profesional, INCE.
- De la Orden, A. Gaviria, J.L. Fuentes, A. y Lázaro, A. (1994) PONENCIA III. Modelos de construcción y validación de instrumentos diagnósticos *Revista de Investigación Educativa, 23*, 129-178.
- Ebel, R.L. (1962): Content standard test scores. *Educational and Psychological Measurement, 22*, 15-25.
- Ebel, R.L. (1972): *Essentials of educational measurement*. Englewood Cliffs, NJ: Prentice-Hall.
- Faggen, J. (1994): *Setting standards for constructed response tests: An overview*. Princeton: NJ. Educational Testing Service.
- Ferrara, S., Perie, M. y Johnson, E. (2002, April): *Setting performance Standard: The item descriptor (ID) matching procedure*. Paper presented at the annual meeting of the American Educational Research Association, New Orleans, LA.
- Fitzpatrick, A.R. (1989): Social influences in standard setting: The effects of social interaction on group judgments. *Review of Educational Research, 59*, 315-328.
- Gaviria, J.J. y Tourón, J. (2000): Evaluación de la educación en Navarra (Informe de Evaluación). Pamplona: Consejería de Educación.
- Gaviria, J.L. y Tourón, J. (2000b): Reflexiones en torno a la evaluación de los sistemas educativos: Un concepto dinámico de eficacia. Ponencia presentada en las *Primeras Jornadas de Medición y Evaluación* (Marzo, 2000). Valencia: Universidad de Valencia.

- 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.
- Gross, L.J. (1982): Standards and criteria: A response to Glass criticism of the Nedelky technique. *Journal of Educational Measurement, 19(2)*, 159-162.
- Grosse, M.E. y Wright, B.D. (1986): Setting, evaluating, and maintaining certification standards with the Rasch model, *Evaluation and the Health Professions, 9 (3)*, 267-285.
- Guion, R.M. (1995): Commentary on values and standards in performance assessment. *Educational Measurement: Issues and Practice, 14,* 25-27.
- Hambleton, R.K. (1984): Validating the test scores. En R.A. Berk (Ed.): *A guide to criterion-referenced test construction*. Baltimore: Johns Hopkins University Press.
- Hambleton, R.K. (1998) : Setting performance standards on achievement tests: Meeting the requirements of Title I. In L.N. Hansche (Ed.), *Handbook for the development of performance standards : Meeting the requirements of Title I* (pp. 97-114). Washington, DC : Council of Chief State School Officers.
- Hambleton, R.K. (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). Mahwah, NJ: Erlbaum.
- Hambleton, R.K., Jaeger, R.M., Plake, B.S. y Mills, C.N. (2000a): *Handbook for setting standards on performance assessment. Washington*, DC: Council of Chief State School Officers.
- Hambleton, R.K., Jaeger, R.M., Plake, B.S. y Mills, C.N. (2000b): Setting performance standards on complex educational assessments. *Applied Psychological Measurement, 24 (4)*, 355-366.
- Hambleton, R.K. y 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., Powell, S. y Eignor, D.R. (1979): Issues and methods for standards setting. En R.K. Hambleton y D.R. Eignor (Ed.): *A practitioner's guide to criterion-referenced test development, validation, and test score usage* (Report No. 70): Amherst: Laboratory of Psychometric and Evaluative Research, School of Education, University of Massachusetts.
- Hambleton, R.K. y Slater, S.C. (1997): Reliability of credentialing examinations and the impact of scoring models and standard-setting policies, *Applied Measurement in Education, 10 (1)*, 19-38.
- Hofstee, W.K.B. (1983): The case for compromise in educational selection and grading. In S.B: *Anderson* y 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 standard setting*. Paper presented at the annual meeting of the National Council on Measurement in Education, New Orleans, LA.
- Impara, J.C. y Plake, B.S. (1997): Standard setting: An alternative approach. *Journal of Educational Measurement, 34*, 353-366.
- Impara, J.C. y Plake, B.S. (1998): Teachers' ability to estimate item difficulty: A test of the assumptions in the Angoff standard setting method. *Journal of Educational Measurement, 35 (1)*, 69-81.

- Individuals with Disabilities Education Act. (1997): Public Law 105-17 (20 U.S.C. 1412a, 16-17).
- Jaeger, R.M. (1982): An iterative structured judgment process for establishing standards on competency test: Theory and application. *Educational Evaluation and Policy Analysis. Win 4*, 4 461-475.
- Jaeger, R.M. (1989): Certification of student competence. In R.L. Linn (Ed.), *Educational measurement* (3rd ed., pp. 485-514). New York: Macmillan.
- Jaeger, R.M. (1991): Selection of judges for standard setting. Educational Measurement: *Issues and Practice, 10*, 3-6.
- Jaeger, R.M. (1995): Setting performance standards through two-stage judgmental policy capturing, *Applied Measurement in Education, 8*, 15-40.
- Jaeger, R.M. y Busch (1984): *The effects of a Delphi modification of the Angoff-Jaeger standard-setting procedure on standards recommended for the National Teacher Examinations*. Paper presented at the annual meeting of the American Educational Research Association, New Orleans.
- Jaeger, R.M. y Mills, G.N. (2001): An integrated judgment procedure for setting standards on complex, Large-scale assessments. In G.J. Cizek (Ed), *Setting performance standards: Concepts, Methods, and Perspectives* (pp. 283-312). Mahwah, NJ: Erlbaum.
- Joint Committee on Standards for Educational Evaluation (1981,1994): *Standards for evaluations of educational programs, projects, and materials*. New York: MacGraw-Hill.
- Jornet, J.M. y Suárez, J.M. (1989a): Conceptualización del Dominio Educativo desde la perspectiva integradora en evaluación referida a criterio. *Bordón*, 41, 2, 237-275.
- Jornet, J.M. y Suárez, J.M. (1989b): Revisión de modelos y métodos en la determinación de estándares y en el establecimiento del punto de corte en evaluación referida a criterio (ERC). *Bordón*, 41, 2, 277-301.
- Jornet, J.M. y Suárez, J.M. (Coords.) (1996). Informe de Validación del Modelo de Evaluación EFO. Informe inédito, presentado ante la Consellería de Trabajo y Asuntos Sociales, de la Generalitat Valenciana.
- Jornet, J.M y Suárez, J.M (1996): Pruebas estandarizadas y evaluación del rendimiento: usos y características métricas. *Revista de Investigación Educativa, 14*, (2), 141-163.
- Jornet, J.M.; Suárez, J.M.; González Such, J. y Belloch, C. (1997): Estrategias de elaboración de pruebas criteriales en Educación Superior, en C. Martínez Mediano (Coord): *Encuentros en la Facultad de Educación sobre Evaluación*. Madrid: UNED.
- Kane, M.T. (1994): Validating the performance standards associated with passing scores. *Review of Educational Research, 64* (3), 425-461.
- Kane, M.T. (2001): So much remains the same: Conception and status of validation in setting standards, In G.J. Cizek (Ed.), *Standard performance standards: Concepts, methods, and perspectives* (pp.53-88). Mahwah, NJ: Erlbaum.
- Kingston, N.M., Kahl, S.R., Sweeney, K., y 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. 219-248). Mahwah, NJ: Erlbaum.
- Lewis, D.M., Mitzel, H.C. y Green, D.R. (1996, June): Standard setting: A book-mark approach. In D.R. Green (Chair), *IRT-based standard setting procedures utilizing be-*

havioural anchoring. Symposium conducted at the Council of Chief State School Officers National Conference on Large-Scale Assessment, Phoenix, AZ.

- Lewis, D.M., Mitzel, H.C. y Green, D.R. y Patz, R.J. (1999): *The bookmark standard setting procedure*. Monterey, CA: McGraw-Hill.
- Linn, R.L. (1978): Demands, cautions, and suggestions for setting standards. *Journal of Educational Measurement, 15* (4), 301-309.
- Linn, R.L. (1994): *The likely impact of performance standards as a function of uses: From rhetoric to sanctions*. Paper presented at the Joint Conference on Standard Setting for Large-Scale Assessments, Washington, DC.
- Livingston, S.A. (1982): Comment on Rowley's paper, Historical antecedents of the standard-setting debate: An inside account of the minimal-beardedness controversy. *Journal of Educational Measurement, 19* (3), 229.
- Livingston, S. A. y Zieky, M. J. (1982): *Passing scores: A manual for setting standards of performance on educational and occupational tests*. Princeton, NJ: Educational Testing Service.
- Livingston, S.A. y Wingersky M.S. (1979): Assessing the reliability of tests used to make pass/fail decisions. *Journal of Educational Measurement. 16*, 247-260.
- Livingston, S.A. y Zieky, M.J. (1982): *Passing scores*. Princeton, NJ: Educational Testing Service.
- Livingston, S.A. y Zieky, M.J. (1982): *Passing Scores: A Manual for setting standards of performance on educational and occupational tests*. Princeton, NJ: Educational Testing Service.
- Loomis, S.C. y Bourque, M.L. (2001): From tradition to innovation: Standard setting on the National Assessment of Educational Progress. In G.J. Cizek (Ed.). *Setting performance standards: Concepts, methods, and perspectives* (pp. 175-218). Mahwah, NJ: Erlbaum
- Madaus, G. F. (1988). The influence of testing on the curriculo. In L. N. Tanner (Ed.) Critical Issues in curriculo. *Eighty-seventh Yearbook on the National Society for Study of Education* (pp. 83-121). Chicago, IL: University of Chicago Press.
- Martínez Rizo, F., Backhoff, E., Castañeda, S., De la Orden, A., Schmelkes, S., Solano-Flores, G., Tristán, A. y Vidal, R. (2000): *Estándares de calidad para instrumentos de evaluación educativa. México*: Ceneval.
- Mehrens, W.A. y Cizek, G.J. (2001): Standard setting and the public good: Benefits accrued and anticipated. In G.J. Cizek (Ed.), *Setting performance standards: Concepts, methods, and perspectives* (pp. 477-485). Mahwah, NJ: Erlbaum.
- Meskauskas, J.A. (1976): Evaluation models for criterion-referenced testing: views regarding mastery and standard-setting. *Review of Educational Research. 46*, 1, 133-158.
- Messick, S. (1975): Historical antecedents of the standard-setting debate: An inside account of the minimal-beardedness controversy. *Journal of Educational Measurement. 19* (2): 87-95.
- Messick, S. (1975): The standard problem: meaning and values in measurement and evalua¬tion. *American Psychologist, 30*, 955-966.
- Messick, S. (1980): Test validity and the ethics of assessment. *American Psychologist, 35*, 1012-1027.
- Messik, S. (1989): Validity. In R.L. Linn (Ed.). *Educational measurement* (3rd ed., pp. 13-104). New York: Macmillan.

- Mitzel, H.C., Lewis, D.M., Patz, R.J. y Green, D.R. (2001): The Bookmark procedure: Psychological perspectives. In G.J. Cizek (Ed.), *Setting performance standards: Concepts, methods, and perspectives* (pp. 249-281). Mahwah, NJ: Erlbaum.
- Muraki, E. (1992): A generalized partial credit model: *Application of an EM algorithm. Applied Psychological Measurement, 16*, 159-176.
- Nedelsky, L. (1954): Absolute grading standards for objective tests. *Educational and Psy¬chological Measurement, 14* (1), 3-19.
- No Child Left Behind Act. (2001): Public Law 107-110 (20 U.S.C. 6311).
- Norcini, J.J., Lipner, R.S., Langdon, L.O. y Strecker, C.A. (1987): A comparison of three variations on a standard-setting method. *Journal of Educational Measurement, 24*, 56-64.
- Pajares, R., Sanz, A., y Rico, L. (2004): *Aproximación a un modelo de evaluación: el proyecto PISA 2000*. Madrid: INECSE.
- Perales, M.J. (2000). Enfoques de evaluación de la Formación Ocupacional y Continua. Estudio de validación de un modelo. Tesis Doctoral. Universitat de València.
- Phillips, S.E. (2001): Legal issues in standard setting for k-12 programs. In G.J. Cizek (Ed.), *Setting performance standards: Concepts, Methods, and Perspectives* (pp. 411-426). Mahwah, NJ: Erlbaum.
- Pitoniak, M.J. (2003): *Standard setting methods for complex licensure examinations. Unpublished doctoral dissertation*. Amherst: University of Massachusetts.
- Plake, B.S. y Hambleton, R.K. (2001): The analytic judgment method for setting standards on complex performance assessments. In G.J. Cizek (Ed.), *Setting performance standards: Concepts, methods, and perspectives* (pp. 283-312). Mahwah, NJ: Erlbaum.
- Plake, B.S.; Melican, G.J. y Milis, C.N. (1991): Factors influencing intrajudge consistency during standard-setting, *Educational Measurement: Issues and Practice, 10* (2), 15-16.
- Putnam, S.E., Pence, P. y Jaeger, R.M. (1995): A multi-stage dominant profile method for setting standards on complex performance assessments. *Applied Measurement in Education, 8 (1)*, 57-83.
- Raymond, M.R. y Reid, J.B. (2001): Who made thee a judge? Selecting and training participants for standard setting. In G.J. Cizek (Ed.), *Setting performance standards: Concepts, methods, and perspectives* (pp. 119-157). Mahwah, NJ: Erlbaum.
- Reckase, M.D. (2001): Innovative methods for helping standard-setting participants to perform their task. The role of feedback regarding consistency, accuracy, and impact. In G.J. Cizek (Ed.), *Setting performance standards: Concepts, methods, and perspectives* (pp. 159-174). Mahwah, NJ: Erlbaum.
- Reid, J.B. (1991): Training judges to generate standard-setting data, *Educational Measurement: Issues and Practice, 10 (2)*, 11-14.
- Rowley, G.L. (1982): Historical antecedents of the standard-setting debate: An inside account of the minimal-beardedness controversy. *Journal of Educational Measurement*. Sum. Vol. 19(2): 87-95.
- Schagen, I. y Bradshaw, J. (2003, September): *Modelling item difficulty for Bookmark standard setting*. Paper presented at the annual meeting of the British Educational Research Association, Edinburgh.
- 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*. Baltimore: Johns Hopkins University Press.
- Shepard, L.A., Glaser, R., Linn, R. y Bohmstedt, G. (1993): *Setting performance standards for achievement tests*. Standford, CA: National Academy of Education.
- Sireci, S.G. (2001): Standard setting using cluster analysis. In G.J. Cizek (Ed), *Setting performance standards: Concepts, Methods, and Perspectives* (pp. 339-354). Mahwah, NJ: Erlbaum.
- Talente, G., Haist, S. y Wilson, J. (2003): A model for setting performance standards for standardized patient examinations. E*valuation and the Health Professions, 26 (4)*, 427-446.
- Thurlow, M.L. y Ysseldyke, J.E. (2001): Standard-setting challenges for special populations. In G.J. Cizek (Ed), *Setting performance standards: Concepts, Methods, and Perspectives* (pp. 387-410). Mahwah, NJ: Erlbaum.
- Van der Linden, W.J. (1982): A latent trait method for determining the intrajudge inconsistency in the Angoff and Nedelsky techniques of setting standards. *Journal of Educational Measurement, 19*, 295-308.
- Wang, N. (2003): Use of the Rasch IRT model in standard setting: An item mapping method. *Journal of Educational Measurement, 40*, 231-253.
- Wright, B.D. y Masters, G.N. (1982): *Rating scale analysis*. Chicago: MESA.
- Wright, B.D. y Stone, M.H. (1979): *Best test design. Chicago*: MESA.
- Ziecky, M.J. (1995): A historical perspective on setting standards. *In Proceedings of Joint Conference on Standard Setting for Large-Scale Assessments*. (pp. 1-38). Washington, DC: National Assessment Governing Board and National Center for Education Statistics.
- Ziecky, M.J. (2001): So much has changed: How the setting of cutscores has evolved since the 1980´s. In G.J. Cizek (Ed), *Setting performance standards: Concepts, methods, and perspectives* (pp.19-52). Mahwah, NJ: Erlbaum.
- Zieky, M.J. y Livingston, S. A. (1977): *Manual for setting standards on the Basic Skills*

Instituto Nacional para la Evaluación de la Educación
José Ma. Velasco No. 101, Col. San José Insurgentes, Deleg. Benito Juárez
C.P. 03900, México, D.F.