Modelo para la determinación de Niveles de Logro y Puntos de Corte de los Exámenes de la Calidad y el Logro Educativos (Excale)

Jesús M. Jornet Meliá* y Eduardo Backhoff Escudero**

CUADERNO No. 30

Instituto Nacional para la Evaluación de la Educación

COLECCIÓN CUADERNOS DE INVESTIGACIÓN

ISSN 1665-9457

Modelo para la determinación de Niveles de Logro y Puntos de Corte de los Exámenes de la Calidad y el Logro Educativos (Excale)

Jesús M. Jornet Meliá* y Eduardo Backhoff Escudero**

CUADERNO No. 30

Colección Cuadernos de Investigación

ISSN 1665-9457

Este texto puede consultarse en: www.inee.edu.mx

*Universidad de Valencia, España. **Instituto Nacional para la Evaluación de la Educación, México

MÉXICO, MARZO, 2008

ÍNDICE

1. Presentación 5
2. Notas acerca del concepto de estándar 6
3. Desarrollo de los métodos para determinar estándares 7
4. Elección del método para la determinación de Niveles de Logro (NL) 10
4.1. Determinación de Niveles de Logro 11
4.2. Determinación de Puntos de Corte 18
5. Validación de los Niveles de Logro 28
6. Conclusiones y recomendaciones de mejora del modelo 35
7. Bibliografía. 38

ÍNDICE DE TABLAS

Tabla I. Ejemplos de etiquetas o nombres de Niveles de Logro 7
Tabla II. Niveles de Logro genéricos para los Excale 11
Tabla III. Modelo del INEE para la determinación de NL 12
Tabla IV. Momentos, actividades y productos del Comité 1, momento 1 14
Tabla V. Ejemplo de un descriptor elaborado por el Comité 1 14
Tabla VI. Estructura de organización de sesiones de trabajo de los comités 1 y 2 15
Tabla VII. Verbos asignados a niveles de rendimiento de acuerdo a su complejidad cognitiva 16
Tabla VIII. Esquematización de niveles y puntuaciones de corte θ_1 20
Tabla IX. Muestra de informaciones para la retroalimentación del Comité 2 25
Tabla X. Síntesis de informes a emitir y unidades de trabajo encargadas 28
Tabla XI. Síntesis de indicadores y fuentes de información 29
Tabla XII. Síntesis de instrumentos para evaluar el proceso de NL y PC 31
Tabla XIII. Ejemplo de síntesis de resultados del cuestionario 1 33
Tabla XIV. Ejemplo de síntesis de resultados del cuestionario 2.1 33
Tabla XV. Ejemplo de síntesis de resultados del cuestionario 2.2 33

ÍNDICE DE FIGURAS

Figura 1. Muestra de retícula de la asignatura Matemáticas (3º de secundaria) Encarte anexo
Figura 2. Ejemplo de un ítem del Cuaderno de Reactivos Ordenados (CRO): 3º de primaria, Matemáticas 18
Figura 3. Ilustración de CRO y reactivos marcadores 19
Figura 4 . Sesiones de juicio del Comité 2 21
Figura 5. Distribución de los jueces y reactivos (para PC1 sugerida) 26
Figura 6. Distribución de estudiantes para el PC1 sugerido 27
Figura 7. Ejemplo de resultados de indicadores de calidad del proceso de identificación de PC: Excale -09 / Español 34
Figura 8. Ejemplo de distribución de PC identificados en tres momentos: Excale-06 / Matemáticas 35

ÍNDICE DE RECUADROS

Recuadro 1. Protocolo para la formación del Comité 1 15
Recuadro 2. Protocolo de trabajo del Comité 1. Momento 1 16
Recuadro 3. Protocolo de trabajo del Comité 1. Momento 2 17
Recuadro 4. Protocolo para la formación del Comité 2 22
Recuadro 5. Protocolo para el formato de juicio 23
Recuadro 6. Protocolo para la retroalimentación 25
Recuadro 7. Protocolo para la selección de los PC 27

Presentación

La determinación de los Niveles de Logro (NL) o de estándares de ejecución de los Exámenes de la Calidad y el Logro Educativos (Excale), del Instituto nacional para la Evaluación de la Educación (INEE), constituye un proceso complejo —propio de la construcción de pruebas de gran escala— y que implica elementos de validación. Las estrategias y procedimientos que se utilizan para lograr un adecuado sistema de interpretación de resultados de las pruebas nacionales se basan, por lo general, en métodos que han sido probados por diversas agencias evaluadoras e instituciones académicas de prestigio internacional.

Una de las claves de validez de la determinación de estándares de interpretación de los resultados de las pruebas educativas es la transparencia del proceso adoptado, tal como lo señala el estándar 4.19 de la APA, AERA y NCME (1999), que dice: "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). La información recabada en este tipo de procesos debe ser la primera garantía acerca de la calidad de los NL identificados.[1]

El proceso para la determinación de NL que ha realizado el INEE parte de su Plan General de Evaluación del Aprendizaje (Backhoff y Díaz, 2005), en el que se describen las siete fases y 16 etapas del proceso de diseño, construcción y validación de los distintos Excale que utiliza la institución para evaluar los aprendizajes de los estudiantes de educación básica de México. La determinación de los NL representa un punto de gran importancia en el diseño de estos instrumentos, pues aporta el sistema de interpretación de los resultados de la evaluación educativa que se realiza para dar a conocer los resultados educativos en el país.

Este proceso puede sustentarse en diversas opciones metodológicas. De hecho, los métodos propuestos para este propósito son innumerables y componen un *corpus* científico importante de la medición educativa. Por ello, se han realizado dos grandes tareas: 1) la selección del método para la determinación de NL y 2) el establecimiento de los estándares para su interpretación.

En este cuaderno técnico se sintetiza la información relativa al modelo diseñado y utilizado para la determinación de los NL de los diversos Excale.

La determinación del sistema de interpretación de las pruebas es un aspecto delicado, pues supone definir con claridad los criterios que se deberán utilizar para valorar la calidad del aprendizaje de los estudiantes del Sistema Educativo Nacional (SEN). Desde esta perspectiva, debe tenerse en cuenta que los Excale fueron diseñados para evaluar los aprendizajes que se definen en el currículo mexicano. Sin embargo, sus resultados podrían ser analizados únicamente desde una óptica normativa, es decir, ordenando a los alumnos en función de los resultados en las pruebas; condición que no constituye en sí misma una interpretación acerca de lo que se conoce y se domina del currículo. Para poder interpretar el nivel de dominio curricular que tiene un estudiante, se requieren establecer criterios claros y suficientes que marquen las habilidades y conocimientos que tiene el alumno en relación con los dominios evaluados. El proceso para determinar estos NL involucra a una gran cantidad de especialistas en educación (expertos en currículo, investigadores educativos, autores de libros de texto y profesores frente a grupo) que trabajan para ello en forma colegiada. Los NL se pueden entender como el compo-

[1] En este sentido, el INEE desarrolló un informe global (Jornet y Backhoff, 2006) falta título en el que se rinden cuentas sobre el proceso para determinar los NL de los Excale-06 (sexto de primaria) y de los Excale-09 (tercero de secundaria) en las asignaturas de Español y Matemáticas.

nente *criterial*, que requieren los Excale para hacer una interpretación correcta de los resultados.

El sistema de interpretación de los Excale constituye uno de los elementos de su validez. La utilidad de la evaluación se basa en una información válida y confiable, la cual represente de forma adecuada los resultados de aprendizaje en un sistema educativo, y permita establecer pautas para la mejora del mismo. Por ello, la determinación de NL constituye un elemento clave para la validez de los Excale.

En el documento *Acerca de la Validez de los Exámenes de Calidad y Logro Educativos (Excale)* (Ruiz-Primo, Jornet y Backhoff 2006) buena parte de los elementos de investigación a considerar para la validación de estos exámenes se refieren a la interpretación de la prueba, y entre ellos se identifican dos grandes conjuntos de acciones: a) la determinación de NL y puntuaciones de corte, y b) la validación de los mismos, basada en la acumulación de diversos tipos de evidencias. Ambos aspectos se han tenido en cuenta en el diseño y desarrollo del modelo metodológico del INEE para la determinación de NL de los Excale que comentamos a continuación.

Notas acerca del concepto de estándar

El término estándar se ha utilizado ampliamente en la literatura de la medición, evaluación e investigación psicológica y educativa, bajo dos grandes acepciones:

- *Normas y procedimientos con las cuales juzgar la calidad de las evaluaciones* (por ejemplo, los estándares establecidos por el Joint Committee on Standards for Educational Evaluation, 1981,1994).
- *Criterios y/o normas para la interpretación de las puntuaciones de los tests psicológicos y/o pruebas de rendimiento educativo*[2].

En el caso de los Excale, nuestro trabajo se centra sobre la segunda acepción aplicada a las pruebas de rendimiento académico.

Por otra parte, bajo el concepto de estándares, como sistema organizado de criterios y/o normas de interpretación de las puntuaciones de las pruebas, distintos autores utilizan como sinónimos los términos *estándares* y *puntuaciones de corte*. Sin embargo, por la confusión que esta práctica puede ocasionar, es importante diferenciar ambos términos (ver a: Van der Linden, 1981; Jornet, 1987; Jornet y Suárez, 1989; Kane, 1994; Cizek, 2001; Cizek, Bunch y Koons, 2004).

De esta forma, es conveniente reservar el término *estándar* para hacer referencia al sistema de criterios de interpretación de resultados de pruebas, es decir la definición teórica de los NL; mientras que el término puntuación de corte (PC) se debe utilizar para indicar la puntuación en la prueba que sirve para diferenciar a los alumnos que se encuentran en uno u otro NL. Como señalaba Van der Linden (1981), el término estándar se debe utilizar para la concepción de los NL en una escala de *puntuaciones verdaderas*, mientras que el término PC representa la diferenciación entre niveles en la escala de *puntajes observados*.

En esta misma línea, Kane (1994) señala que:

> …" es útil marcar una diferencia entre la puntuación de 'pase', definida como un punto en una escala de puntuaciones, y el estándar de rendimiento, definido como el nivel… mínimo suficiente de rendimiento para algún propósito… El estándar de rendimiento es la versión conceptual del nivel deseado de competencia, y la puntuación de 'pase' es la versión operativa.")

En cualquier caso, los términos estándares y PC aluden a dos aspectos de un mismo proceso. En la definición de los estándares se identifican cuatro elementos con claridad:

1. **Categorías relativas a los NL**. El número de categorías que se utilizan en una prueba de aprendizaje pueden ser variables; comúnmente, se suele utilizar de tres a seis categorías. Las categorías se describen con nombres o etiquetas que aluden al nivel de dominio sobre un área de competencia en particular (ver ejemplos de la tabla I); asimismo, las categorías se pueden identificar o etiquetar con números (nivel 1, 2,…, n) en vez de nombres.
2. **Descriptores de los NL**. Para identificar cada uno de los NL de una prueba, se utilizan descripciones sintéticas, o descriptores, que reflejan de forma global el tipo de aprendizaje adquirido por los estudiantes que se pueden clasificar en cada categoría. Además, estas descripciones se complementan con ejemplos del tipo de ejecuciones académicas que son capaces de realizar los alumnos ubicados en cada NL, además de una muestra del reactivo[3] con el que se evalúa la ejecución.
3. **Puntuaciones de corte (PC)**. Se refieren a las puntuaciones que en la prueba sirven para diferenciar o distinguir cada uno de los NL establecidos.

[2]Reservamos el término *test* para los instrumentos de medición de rasgos psicológicos y el término *prueba* para aquéllos destinados a la evaluación de rendimiento o logro académico.

[3]Los términos reactivo, ítem o pregunta se utilizan aquí como sinónimos.

Tabla I. Ejemplos de etiquetas o nombres de Niveles de Logro

Etiquetas	Origen
Básico, Competente, Avanzado	National Assessment of Educational Progress
En camino, Progresando, Cerca de la competencia, Competente, Avanzado	Terranova (2ª. ed.) (CTB / McGraw Hill)
Limitado, Básico, Competente, Acelerado, Avanzado	Pruebas de rendimiento del Estado de Ohio
Muy por debajo del nivel básico, Debajo del nivel básico, Básico, Competente, Avanzado	Pruebas del Estado de California
No llega al nivel usual, Llega al nivel usual, Rendimiento destacado	Estado de Texas, Estándares de valoración de Texas de conocimientos y destrezas
Sin etiquetas, identificación mediante la cualificación de la escala numérica	INCE[4]: Diagnóstico del sistema educativo español, 1998
Cinco niveles, sin etiquetas	Proyecto PISA (2000): Lectura
Escala de habilidad: Máximo, Medio, Mínimo	Proyecto PISA (2000): Matemáticas y Ciencias

Fuente: Adaptado de Cizeck, Bunch y Koons, 2004.

4. **Items característicos**. Se refieren a los reactivos que son capaces de responder correctamente los estudiantes que se ubican en un determinado NL, de forma diferencial respecto de los demás niveles.

Los elementos que se refieren a los puntos 1 y 2 (etiquetas y descriptores) se suelen establecer por procedimientos de juicio, mientras que para los elementos 3 y 4 normalmente se tiene en cuenta el funcionamiento empírico de la prueba. Sobre ello volveremos después al revisar los tipos de métodos de determinación de estándares.

El proceso de determinación de NL incluirá por lo tanto todos los elementos mencionados, de forma que haga referencia a un conjunto de acciones que permitan definir desde un sistema de categorías, su descripción y la identificación de PC e items característicos de cada NL en los Excale. En este sentido, podríamos definir este proceso como de carácter político-técnico, dado que implica que se tomen decisiones en estas dos dimensiones. Como señala Cizek (2001b):

> "El establecimiento de estándares es quizás la rama de la psicometría que mezcla —más que cualquier otra— ingredientes artísticos, políticos y culturales en la preparación de sus productos " (p.5).

En síntesis, la finalidad del proceso de determinación de NL es poder aportar un sistema de interpretación de puntuaciones de las pruebas de rendimiento que esté al servicio de la toma de decisiones y de la comunicación de resultados. Por lo anterior, es altamente deseable que los NL de las pruebas de aprendizaje tengan un alto grado de confiabilidad y validez, y que sean parte integral del proceso de diseño, construcción y validación de cualquier prueba de aprendizaje que se utilice a nivel nacional.

Desarrollo de los métodos para determinar estándares

En este apartado, nos centraremos en describir algunas características y problemas de los métodos comúnmente utilizados para la determinación de NL; asimismo, se expondrán los argumentos de mayor peso que han ido guiando su desarrollo o evolución, con la finalidad de justificar las características del modelo utilizado por el INEE para determinar los NL de los Excale (tema central del próximo apartado).

Los procedimientos para la determinación de estándares se plantean y desarrollan en el ámbito de las *pruebas criteriales*, desde la década de los años sesenta[4] del siglo pasado. El problema que se plantea en ese momento es dar una respuesta adecuada al tipo de decisiones que deben tomarse a partir de los resultados de las pruebas de logro académico. Así, el planteamiento inicial era desarrollar métodos que permitieran aportar una valoración *absoluta* de calidad de las ejecuciones de los estudiantes en las pruebas educativas; dado que, hasta ese entonces, la interpretación de las puntuaciones se basaba en

[4]En 1963 Robert Glaser publicó su artículo *Instructional technology and the measurement of learning outcomes: some questions,* en el que se plantean las bases de desarrollo de las pruebas criteriales.

las normas de referencia del grupo de estudiantes que respondía la prueba. Para algunos autores esta forma de interpretar los resultados de las pruebas de rendimiento académico constituía una gran limitación de base para el tipo de juicios que era necesario tomar en el ámbito educativo[5]. En la lógica subyacente a este tipo de propuestas se identifica el hecho que cualquier profesor frente a grupo realiza este tipo de valoraciones para promover, o no, el aprendizaje de los estudiantes en los diversos programas educativos. El planteamiento, en todo caso, era hacer objetivo este tipo de juicios.

En ese marco, y hasta la década de los años ochenta, el problema sobre el que se inicia el desarrollo de esta área es el establecimiento de puntuaciones destinadas a identificar a los estudiantes en dos grupos: los que dominan el contenido educativo y los que no lo dominan, decisión que responde a una situación normal en el proceso educativo. A pesar de que es muy común adoptar este tipo de decisiones en el ámbito escolar, los problemas de medición implicados en las mismas son bastante complejos, pudiéndose equiparar a los que se suscitan al tratar de medir un constructo psicológico no observable.

A partir de la década de los noventa, el énfasis en la investigación evaluativa se centra en el desarrollo de esquemas de valoración politómicos; es decir, aquellos que se basan en el establecimiento de estándares de más de dos categorías de ejecución. Lo anterior se produce como consecuencia de las mejoras que permiten los modelos de la Teoría de Respuesta al Item (IRT, por sus siglas en inglés), las innovaciones en el desarrollo de las pruebas de aprendizaje de gran escala y los estudios de la calidad de los sistemas educativos nacionales e internacionales.

Es el inicio de lo que hoy se conoce como *pruebas referidas a estándares*. Por ejemplo, el programa norteamericano *National Assessment for Educational Program* (NAEP) fue uno de los primeros en expresar los niveles de rendimiento de los estudiantes a partir de series graduadas de niveles de ejecución de los alumnos, para lo cual estableció tres NL: Básico, Competente y Avanzado (Cizek, Bunch y Koons, 2004). En España, el estudio *Diagnóstico del Sistema Educativo Español*, realizado por De la Orden y colaboradores (1998), utilizó una escala graduada empírica en la que se identifican los NL a partir de los items característicos de cada uno de ellos. Por su parte, los estudios internacionales de mayor prestigio, como el Programa para la Evaluación Internacional de los Estudiantes (PISA, por sus siglas en inglés) (2000, 2004), también han adoptado sistemas politómicos para comunicar sus resultados.

Los problemas metodológicos del desarrollo de sistemas de interpretación politómicos válidos y confiables han evolucionado también conforme se han refinado los marcos teóricos de su concepción; desde los referidos a la orientación general de este tipo de procesos —como por ejemplo el rol de las tareas de juicio de expertos frente a la información empírica— hasta los problemas muy específicos que ponen de manifiesto la madurez de este ámbito de la investigación evaluativa (Ziecky, 1995, 2001).

El problema sobre el que se centró el debate inicial acerca de los métodos para la determinación de estándares, fue el planteado por Glass (l978) sobre la arbitrariedad de los procedimientos para la determinación de los PC. Nuestra posición, aun reconociendo las limitaciones reflejadas por Glass (1978), se identifica con la concepción que en la actualidad ha venido prevaleciendo, que resalta la necesidad, utilidad y valor de las tareas de juicio de expertos en estos procedimientos, y que en su momento defendieron, frente a Glass, autores como Popham, (1978ª), Block (1978), Hambleton (1978) Shepard (1980ª, 1984) y Berk (1980, 1984). No obstante, el reconocimiento de que el juicio humano es fundamental en los procedimientos orientados a la adopción de una decisión final sobre la suficiencia/insuficiencia del aprendizaje del estudiante, no supone la solución definitiva de todos los problemas involucrados; únicamente pone énfasis en el reconocimiento de las limitaciones básicas de este tipo de acercamientos metodológicos.

Entre los métodos que se fueron desarrollando destinados a la identificación de PC se pueden identificar tanto aquellos que se basan exclusivamente en el juicio humano (ya sea sobre los reactivos o sobre los estudiantes), como aquellos que combinan el juicio humano con elementos empíricos (resultados de las pruebas). Desde sus primeros años de desarrollo, se produjo una gran proliferación de métodos, si bien buena parte de ellos eran adaptaciones y/o extensiones de otros. Hasta el momento, se han presentado diversos sistemas de clasificación respecto a estos métodos y procedimientos (véase a: Meskauskas, 1976; Glass, l978; Shepard, 1980, 1984;

[5]Lo que se necesita saber para decidir acerca de si un alumno domina lo suficiente un área o dominio educativo es poder valorar si "sabe o no sabe", no si "sabe más o menos que sus compañeros". Es decir, es necesario disponer de valoraciones absolutas acerca de la calidad de los aprendizajes de los estudiantes; no valoraciones de carácter relativo, como las propias de los tests psicométricos clásicos.

Hambleton, 1980; Berk, 1986; Jornet, 1987; Jornet y Suárez, 1989; Cizeck, 1996[a]; Hambleton et al., 2000[a] y [b]; Cizeck, Bunch y Koons, 2004), de los cuales podemos destacar la siguiente clasificación:

- **Métodos de juicio.** Aquéllos que basan el establecimiento de los estándares en el juicio que realizan expertos acerca de los items, los sujetos o las tareas[6].
- **Métodos empíricos**. Éstos incluyen a los que priorizan los criterios estadísticos para apoyar la calidad de la decisión; entre ellos se pueden clasificar los modelos de estado[7] y los continuos basados en la teoría de la decisión[8].
- **Métodos mixtos.** Son aquellos que partiendo de valoraciones basadas en juicio de expertos, ajustan la identificación de las puntuaciones de corte considerando elementos empíricos del funcionamiento de las pruebas; entre ellos se pueden identificar los métodos de compromiso[9] y los de correspondencia de items[10].

Dada la enorme oferta metodológica existente, un problema central en el establecimiento de estándares, es saber qué método elegir. Como se señala en los *Estándares para la evaluación educactiva y psicológica* (AERA, APA, NCME, 1999) no hay un único método para determinar los PC para todas las pruebas y para todos los propósitos, ni puede haber un único conjunto de procedimientos para establecer su justificación. Junto a este problema hay una realidad que tranquiliza: la evolución de los métodos, así como los estudios comparativos realizados al respecto, ofrecen al menos criterios claros que pueden ayudar a centrar el método a elegir, y que han sido expuestos en un trabajo anterior (Jornet y Perales, 2001).

Además del método específico que se utilice para la determinación de estándares, es preciso seguir unas etapas generales para desarrollar este proceso. Hambleton (1998, 2001) presentó una síntesis de los pasos a seguir, la cual resumimos a continuación:

1. Seleccionar un comité de expertos grande y representativo, como base de la validez y confiabilidad de los estándares.
2. Elegir el método de determinación de estándares; preparar materiales de formación y el programa de reuniones para la determinación de estándares.
3. Preparar las descripciones de las categorías de rendimiento.
4. Formar a los participantes en el uso del método de determinación de estándares.
5. Recopilar clasificaciones de items y otras valoraciones de los participantes y producir información descriptiva, cuyo propósito es retroalimentar a los participantes.
6. Facilitar la discusión entre los participantes de la información descriptiva/resumen inicial.
7. Realizar una segunda sesión de clasificaciones/valoraciones; compilar la información y facilitar la discusión como en los pasos 5 y 6.
8. Dar una oportunidad final a los participantes de examinar la información y llegar a los estándares finales de rendimiento recomendados.
9. Llevar a cabo una evaluación del proceso de determinación de estándares, recogiendo información sobre la confianza de los participantes en el proceso y los estándares de rendimiento resultantes.
10. Reunir la documentación del proceso de determinación de estándares y cualquier otra evidencia de la validez de los estándares de rendimiento resultantes.

[6]Entre los denominados *métodos de juicio*, destacan los métodos basados en el juicio sobre los items, como los de Nedelsky (1954), Angoff (1971), Jaeger (1978), o Ebel (1979). Otro grupo de métodos que ha tenido también buena acogida y trascendencia ha sido el de métodos basados en el juicio sobre sujetos. Sistematizados por Livingston y Zieky (1982), han tenido una amplia aplicación y uso. Una evolución metodológica que podríamos situar entre los dos conjuntos de métodos mencionados son los de *juicio sobre tareas*, también denominados *métodos holistas* (Cizeck, Bunch y Koons, 2004). En esta categoría también se incluyen diversos métodos, como el de *juicio analítico* de Plake y Hambleton (2001), el método de *selección de trabajos* de Loomnis y Bourque (2001), el método *The body of work method* (método del cuerpo de trabajo) propuesto por Kingston, Kahl, Sweeney y Bay (2001).

[7]Como es el caso de los modelos de estado de Roudabush (1974), o el de Emrick y Adams (1969) y Emrick (1971), revisados por Macready y Dayton (1980).

[8]En su momento dieron origen a diversos procedimientos, que se diferenciaban básicamente en la consideración del error, tomando como referencia diferentes funciones de pérdida (*umbral* –Hambleton y Novick, 1973; Novick et al., 1973; Swaminathan et al. 1975-, *lineal* –Huynh, 1976; Van der Linden y Mellenberg, 1977; Mellenberg y Van der Linden, 1981-, *en ojiva normal* –Novick y Lindley, 1978-, *potencia* –Huynh, 1980-) (Jornet, 1987; Jornet y Suárez, 1989).

[9]Entre ellos, se pueden identificar los métodos de De Gruijter (1980, 1982), Hoffstee (1983) y el de Beuck (1984), descritos por Shepard (1984) y Cizek, (1996a).

[10]Uno de los métodos entre los que actualmente tienen mayor impacto: el *método Bookmark o del marcador*. Presentado por Lewis, Mitzel y Green (1996) y Lewis, Mitzel, Green y Patz (1999), se ha utilizado ampliamente en educación k-12.

Elección del método para la determinación de Niveles de Logro (NL)

Un aspecto básico en la determinación de los NL de los Excale residió en la selección del método a utilizar. La diversidad de métodos propuestos para la determinación de estándares de interpretación e identificación de PC es muy amplia, por lo que el INEE tuvo que valorar las bondades y limitaciones de las diversas opciones metodológicas tomando en cuenta las características propias de los Excale. Un denominador común de estas pruebas es el trabajo colegiado y colaborativo en el diseño, desarrollo y validación de las pruebas de aprendizaje del INEE. Subyacente a este planteamiento, se identifican dos particularidades importantes de mencionar:

- El concepto de evaluación como un proceso multidisciplinario de especialistas en currículo, investigadores en educación, autores de libros de texto, expertos en psicometría y profesores frente a grupo.
- El uso de metodologías cuantitativas y cualitativas complementarias, como el medio más idóneo, confiable y válido para el desarrollo de pruebas de gran escala cuyo propósito es la evaluación de sistemas y subsistemas educativos.

Desde esta posición, la primera tarea del proceso de determinación de NL e identificación de PC consistió en la elección del método. La primera etapa del procedimiento se basó en el desarrollo de un seminario[11] en cual, además de revisar las alternativas metodológicas disponibles, se valoró una propuesta inicial para la determinación de los estándares, se recogieron sugerencias y se ajustó el modelo del INEE. Los productos de este seminario fueron los siguientes:

1. Elección del modelo de determinación NL, como una adaptación del método Bookmark (o "del marcador"), que describiremos en el apartado siguiente.
2. Definición de las categorías y etiquetas generales de los Excale, que también se incluyen en el siguiente apartado.
3. Formación inicial de los responsables de los Excale como coordinadores de los comités para establecer los NL e identificar los PC correspondientes.

El logro educativo que evalúan los Excale representa un constructo de aprendizaje que se sustenta en el contenido curricular mexicano y se asume que es de carácter continuo. De este modo, el supuesto básico es que los Excale actúan como instrumentos que evalúan las competencias escolares de los estudiantes en distintas áreas curriculares. Por ello, la determinación de NL diferenciales que simplifiquen y faciliten la interpretación de los resultados de los Excale debe basarse en la identificación de PC que indican, con una elevada confiabilidad, tipologías diferenciales de la ejecución de los alumnos a lo largo de un continuo de aprendizaje.

Si bien no puede asumirse que haya un solo procedimiento para la determinación de NL y PC que sobresalga sobre los demás por su calidad y pertinencia, sí parecen claros algunos principios básicos que se deben considerar, los cuales ponen de manifiesto las ventajas de los métodos: 1) centrados en los reactivos, 2) de carácter mixto, con juicio de expertos e información empírica y 3) con un componente de retroalimentación sobre los juicios de los expertos.

Asimismo, creemos que hay que asumir el costo de la *arbitrariedad* que preside a toda interpretación de resultados educativos y sociales, lo cual no significa que las decisiones sean caprichosas, sino que se trata de juicios intersubjetivos[12]. En este sentido, el procedimiento para la determinación de NL debe cumplir básicamente los siguientes objetivos de uso:

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 asignaturas evaluadas, por lo que éste 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 mexicana, por lo que deberá atenderse a la distribución y características del comportamiento académico de los alumnos del Sistema Educativo Nacional (SEN) ante los Excale.
3. Que permita validar los NL y PC resultantes a partir de estudios empíricos.

La calidad de los estándares de interpretación de los NL debe ser tal que permita su interpretación para la evaluación actual, así como facilitar interpre-

[11] Al final de este cuaderno se aporta la información acerca de los participantes en el mismo.

[12] El concepto de calidad del aprendizaje, en definitiva, es arbitrario, depende de múltiples factores históricos, sociales y personales. Por ello, si se requiere interpretar qué es un aprendizaje de calidad –en cualquier materia o disciplina– es necesario definirlo de forma operativa.

taciones de carácter longitudinal y transversal de la situación y evolución de los resultados del SEN.

El proceso de determinación de los NL se basa en el trabajo coordinado de dos comités con características distintas (en lo sucesivo Comité 1 y Comité 2) de especialistas en currículo y en investigación educativa, para el primero de ellos, y de profesores en ejercicio, para el segundo. El primero tiene como finalidad la determinación los NL; es decir, de las habilidades y conocimientos característicos de cada asignatura y NL. La tarea del Comité 2 es la identificación de las PC en los Excale, que deben diferenciar a los estudiantes de acuerdo a su nivel de competencias escolares.

Determinación de Niveles de Logro

Como ya se mencionó previamente, la determinación de los NL se basa usualmente en un sistema de tres a seis categorías, y cada una suele estar representada por una etiqueta alusiva al estándar de ejecución del dominio curricular correspondiente. La finalidad de etiquetar o nombrar los NL es disponer de una referencia corta sobre el nivel de dominio que poseen estudiantes, de modo que los especialistas que identifican las habilidades y conocimientos correspondientes a cada NL puedan disponer de un marco conceptual común para todos los Excale. Hay que tener en cuenta que este sistema de categorías de rendimiento escolar también debe servir para la comunicación de los resultados de aprendizaje.

Los criterios que se tuvieron en cuenta para identificar el sistema de etiquetas fueron los siguientes:

1. Simplicidad suficiente para: a) sintetizar la información en un reducido número de categorías y b) 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 de alumnos por su nivel de aprendizaje: 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 dominio muy elevado.
3. Las etiquetas, si bien representan NL, deben ser entendidas como meros identificadores, y de ser posible de forma que no puedan ser interpretadas de forma negativa por la población (por ejemplo, evitar términos como *inferior*, *reprobado*, *superior* equivalentes).
4. Deben evitarse etiquetas ambiguas o que incluyan tecnicismos de difícil comprensión para la sociedad en general, quien 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: a) el funcionamiento de la prueba, pues hay que considerar si la prueba dispone de capacidad suficiente de discriminación para el sistema previsto y b) 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, estableció la categorización y definición de las siguientes etiquetas que se muestran en la tabla II.

Tabla II. Niveles de Logro genéricos para los Excale

Nivel	Descriptor
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.
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.
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.
Por debajo del 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.

El primer comité, que denominamos Comité 1, se encarga de la elaboración de los descriptores de los NL, mientras que el segundo, que denominamos Comité 2, se encarga de identificar los PC o reactivos que sirven de punto de inflexión entre dos niveles de ejecución. Las fases del modelo para determinar los NL y PC de los Excale se sintetizan en la tabla III.

Tabla III. Modelo del INEE para la determinación de NL

Fases/momentos	Responsable(s)	Forma de Trabajo	Productos
Fase 1 Elección del modelo para la determinación de NL de los Excale	• Conductor del seminario • Consejo Técnico del INEE • Dirección de Pruebas y Medición del INEE[1] • Representantes de la SEP	**Seminario** 1. Formación 2. Debate de propuestas 3. Elección del modelo del INEE	• Elección del modelo de determinación de NL • 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 y Medición del INEE		Categorización y definición de etiquetas
Fase 3 Determinación de NL (para cada Excale)			
Momento 0 Elaboración de las especificaciones	Dirección de Pruebas y Medición del INEE		Propuesta de clasificación de especificaciones de subdominios, ordenados por dificultad
Momento 1 Elaboración de elementos genéricos del descriptor	**Comité 1** Comités de descripción de NL. Cada comité está compuesto por: • Coordinador de prueba • Especialistas en currículo y en investigación educativa	1. Formación del comité 2. Grupo de discusión Análisis de los subdominios de cada área evaluada, con asignación de descriptores a cada nivel de logro, para 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 PC (para cada Excale)	**Comité 2** Comités de determinación de PC. Cada comité está compuesto por: • Coordinador de prueba • Cinco docentes en ejercicio	1. Formación 2. Toma de contacto con la prueba 3. Sesiones de juicio 4. Sesión de evaluación del proceso	• Revisión de las descripciones de los NL realizada por el Comité 1 • Identificación de reactivos marcadores y puntuaciones de corte entre categorías o NL, para cada prueba • Valoraciones del proceso – Estudio de Validación
Fase 5 Determinación de NL (para cada Excale)			
Momento 2 Elaboración de las ejemplificaciones de los descriptores de los NL	**Comité 1** Comités de descripción de NL. Cada comité está compuesto por: • Coordinador de prueba • Especialistas en currículo y en investigación educativa	Grupo de discusión	Ejemplificaciones de los descriptores de cada Excale
		Trabajo individual	• Valoración del proceso – Estudio de Validación (Cuestionario 1)

Comité 1

El comité de elaboración de los NL se compone de un número reducido de especialistas y fue independiente del comité que posteriormente trabaja en la identificación de los PC. Estuvo integrado por especialistas en currículo y en investigación educativa, así como por el coordinador(a) de cada prueba, que actúa como conductor(a) del comité. El esquema de trabajo que desarrolla éste se presenta en la tabla III.

Este comité valora 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. De este modo, no se pretende que los participantes en el comité anticipen el comportamiento empírico de la muestra, sino que orienten 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 trata de ajustar las categorías de descripción de los NL, representándolos adecuadamente y aportando muestras de ejecución de los reactivos de cada categoría o nivel de logro.

Este comité trabaja en dos momentos (ver tabla III):

1. *Previo al análisis empírico de resultados*[13], donde se desarrollan los descriptores de los NL con el fin de servir 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

Como se muestra en las tablas IV y V, el procedimiento para su desarrollo es el de panel de discusión, en el cual los miembros del comité llegan a un acuerdo acerca de las categorías de descripción de los niveles, se trata de que lleguen a acuerdos respecto a[14]:

- 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 figura I anexa).
- Una vez que se dispone de la identificación de las puntuaciones de corte que el Comité 2 establece, 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.
- Los elementos de trabajo son tres: 1) descripción genérica de cada nivel, 2) propuesta de clasificación de especificaciones de subdominios, ordenados por dificultad y 3) retícula curricular de la asignatura.

[13]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.

[14]No se trata 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; por el contrario, se trata de que mediante el debate los miembros del comité establezcan acuerdos.

Figura 1. Muestra de retícula de la asignatura Matemáticas (3º de secundaria)

INSTITUTO NACIONAL PARA LA EVALUACIÓN DE LA EDUCACIÓN

DIRECCIÓN DE PRUEBAS Y MEDICIÓN

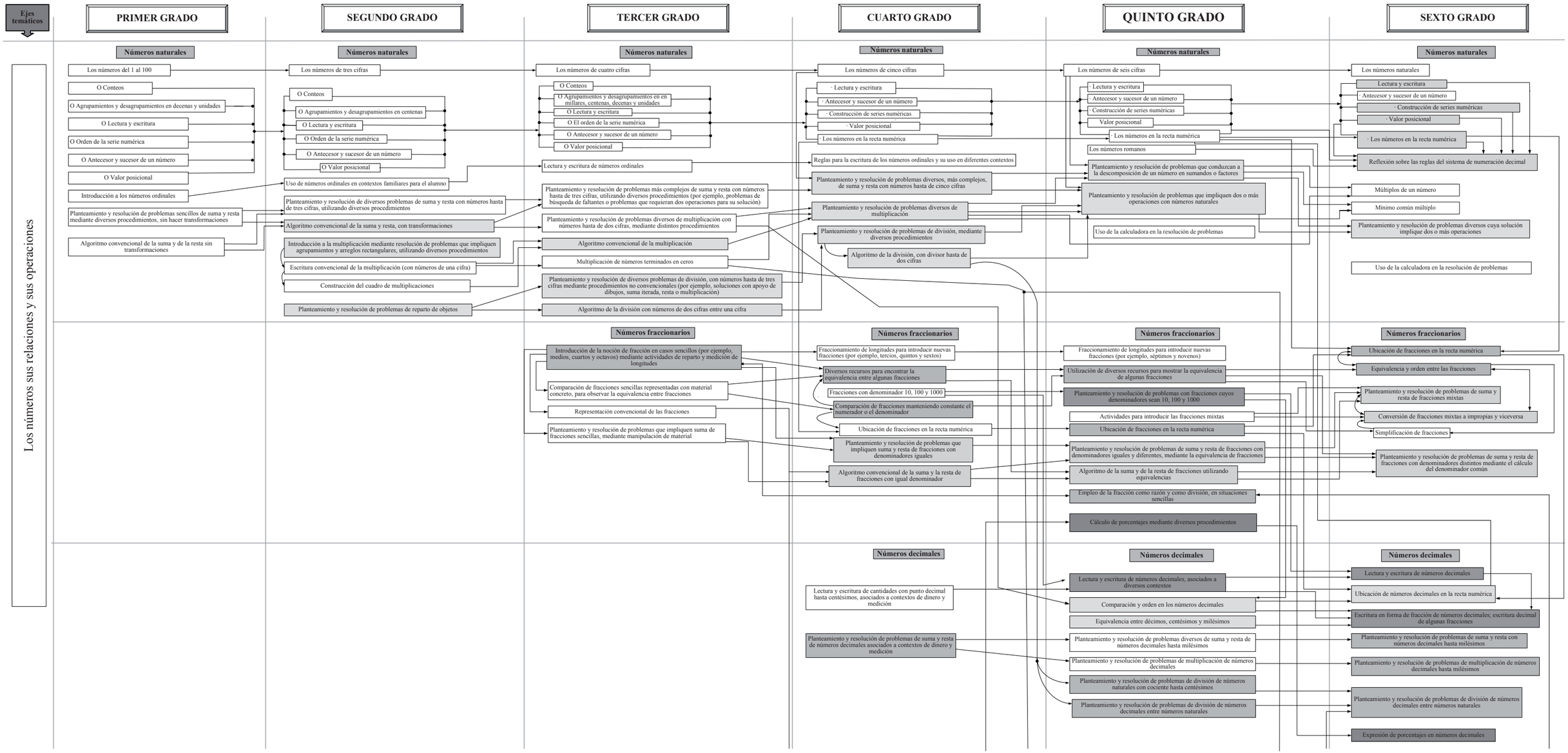


Tabla IV. Momentos, actividades y productos del Comité 1, momento 1

Momentos	Forma de trabajo	Actividades	Productos
1.1	Trabajo grupal	• Primera valoración de la propuesta de etiquetas planteada por la dirección de pruebas del INEE • Explicación global del trabajo de la sesión	
1.2	Trabajo en diadas	• Diada A trabaja subdominio 1 • Diada B trabaja subdominio 2 • Diada … trabaja subdominio … • Diada M trabaja subdominio N	Identificación de especificaciones para los niveles: • Por debajo del básico • Básico • Medio • Avanzado
1.3	Trabajo grupal	• Segunda valoración de la propuesta de etiquetas planteada por la dirección de pruebas del INEE • Valoración del trabajo realizado para cada subdominio • Acuerdos	Considerando todos los subdominios, descripción global de los niveles: • Por debajo del básico • Básico • Medio • Avanzado

Tabla V. Ejemplo de un descriptor elaborado por el Comité 1

Etiqueta del NL	Descriptor	Momentos
Avanzado	Los estudiantes de 4° grado que están en este nivel deben poder generalizar sobre los tópicos en la selección de lecturas y demostrar un conocimiento suficiente acerca de cómo los escritores componen y usan las estrategias literarias. Cuando leen textos apropiados para su nivel, deben poder juzgarlos de forma crítica y, en general, dar respuestas minuciosas que demuestren que han comprendido el texto.	Momento 1: descripción genérica
	Por ejemplo, cuando leen textos literarios los estudiantes deben poder hacer las generalizaciones sobre lo relevante de la historia y prolongar su significado integrando las experiencias personales y las otras interpretaciones con las ideas indicadas por el texto. Deben poder identificar los recursos literarios, como la lengua figurada.	Momento 2: Ejemplo 1
	Cuando leen textos informativos, los alumnos de 4° de nivel Avanzado deben poder explicar el propósito del escritor usando material de soporte del texto. Deben poder hacer juicios críticos sobre la forma y el contenido del texto y explicar sus juicios claramente.	Ejemplo 2

Protocolos de actuación

La organización del trabajo del Comité 1 se especifica mediante protocolos de actuación—o guías de trabajo— en los que se establecen las líneas generales que deben seguirse 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 este comité son los siguientes: 1) protocolo para la formación y 2) protocolo para la elaboración de descriptores.

El tiempo de trabajo que se destina a los comités de tipo 1 es, aproximadamente, 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 VI).

Tabla VI. Estructura de organización de sesiones de trabajo de los comités 1 y 2

Horario	Días				
	1	2	3	4	5
Mañana	Comité 1	Comité 1	Comité 2	Comité 2	Comité 2
Tarde	Comité 1	Comité 2	Comité 2	Comité 2	Comité 1

Recuadro 1

Protocolo para la formación del Comité 1

La formación que se ofrece al Comité 1 para realizar su tarea contempla los siguientes contenidos:

1. Explicar la finalidad y objetivos de la determinación de NL.
2. Mostrar las definiciones realizadas por otras instituciones. En cada caso, se muestran ejemplos de la materia sobre la que debe trabajar cada comité.
3. Presentar la propuesta de etiquetas que utiliza el INEE, el procedimiento por el que se han determinado (propuesta del INEE, criterios que la justifican), así como la definición genérica que corresponde a cada una de ellas.
4. Presentar la forma de trabajo del comité:
   a) Momentos 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: trabaja directamente con las especificaciones utilizadas para construir los reactivos. Para ello:
      - El INEE clasificó los descriptores en subdominios previamente ordenados por dificultad; esa clasificación se aporta como punto de partida para el trabajo del comité
      - Se utiliza la retícula como referencia para identificar la posición en el currículo de cada uno de los descriptores
      - Se presenta al conjunto del comité por parte del coordinador y se conformaron díadas, las cuales trabajan los subdominios completos de cada prueba, identificando los descriptores correspondientes a cada nivel.
      - Desarrollados todos los descriptores por las díadas, se debate cada solución por el conjunto del comité. Una vez que se llega a un acuerdo para cada subdominio, se redacta un párrafo que recoja las características globales de todos ellos para cada nivel. Respecto al nivel *Por debajo del básico*, se entiende que puede llegar a definirse por exclusión, es decir, por no llegar a satisfacer las características del nivel *Básico*, o bien por corresponder a competencias propias de niveles anteriores al trabajado.
5. Síntesis de la tarea a realizar.

Recuadro 2

Protocolo de trabajo del Comité 1

Momento 1 del Comité 1: elaboración de elementos genéricos del descriptor

Se le presentan al Comité 1 los siguientes elementos:

1. Descripción genérica de cada nivel de competencia.
2. Propuesta de clasificación de especificaciones realizada por el INEE.
3. Se forman díadas y se asignaran los subdominios de trabajo a cada una de ellas. En los casos en que esto no sea posible, y con el fin de evitar que se formen las mismas parejas en cada ocasión, las díadas se varían hasta trabajar el conjunto de subdominios (por ejemplo, 1-2, 3-4; 1-3, 2-4).
4. Cada díada debate sobre un subdominio y un nivel. Para ello, se procede del siguiente modo:
    - Se analiza la clasificación desde el nivel *Básico* hasta el *Avanzado*.
    - Concluida la revisión del *Básico*, se analiza la clasificación del nivel *Medio*. Se enfatiza a los miembros del comité que se aseguraren 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 asumió como guía la clasificación de verbos por niveles[1] que se recoge en la tabla VII.
5. Una vez que se llega a un acuerdo en la clasificación de especificaciones, se elabora un párrafo que sintetice de manera global el tipo de competencias, habilidades y/o conocimientos que caracterizan a los estudiantes de cada nivel.
6. Se asume que existe acuerdo cuando hay unanimidad (o asentimiento por parte de todos los miembros del comité). En ningún caso se procede a votar, pero el coordinador del comité debe atender a que no se asuman acuerdos existiendo participantes que estén claramente en desacuerdo; de forma que se mantenga el debate mientras se dan posiciones diferentes, dirigiéndolo permanentemente hasta conseguir el consenso, sin adoptar una actitud directiva que soslaye las posiciones de los participantes.

Tabla VII. Verbos asignados a niveles de rendimiento de acuerdo a su complejidad cognitiva

Niveles de complejidad cognitiva			
Menor (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

Recuadro 2 (continuación)

7. Para proceder al ajuste final de descriptores se atienden 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
    - Asimismo, debe ser comprensible para la mayor parte de personas a las que afecta la evaluación (técnicos y especialistas, profesorado, padres), de forma que facilite la difusión de resultados a la sociedad
    - No se hará referencia a contenidos, sino a competencias, habilidades, destrezas, adquisiciones, conocimientos, maestrías, dominios, logros
    - 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 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

Recuadro 3

Protocolo de trabajo del Comité 1

Momento 2 del Comité 1: elaboración de ejemplos del descriptor

Una vez que se dispone de las puntuaciones de corte identificadas por el Comité 2, el Comité 1 recibe la información acerca de los reactivos que corresponden a cada nivel, así como de sus especificaciones. A partir de esta información:

1. Se compara la definición original que aprobó el Comité 1 con la resultante del Comité 2, de forma que puedan identificarse las discrepancias entre ambas propuestas.
2. Identificadas las diferencias en cada nivel, se valora el grado en que afectan a la redacción del descriptor y se ajusta de forma precisa a la clasificación que emana de la identificación de puntuaciones de corte realizada por el Comité 2.
3. Posteriormente, el Comité 1 redacta las ejemplificaciones (ver tabla IV) correspondientes a cada NL. Para este cometido, se incluye una frase o descripción que ejemplifica las características de los estudiantes del nivel correspondiente en cada una de las competencias del mismo. Para ello, se toma 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 corresponda.
4. Finalmente, se selecciona un reactivo de cada subdominio como muestra del tipo de tareas que pueden realizar los estudiantes pertenecientes a cada nivel. Esta selección es muy cuidadosa, pues se trata de identificar los reactivos que se publicarán en los informes y comunicados del INEE. Para elegirlo, se tiene 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 del PC.
    - Su representatividad respecto a las competencias que caracterizan al nivel.
5. Como en el caso de la redacción de descriptores, se tienen en cuenta las consideraciones realizadas anteriormente en relación al lenguaje a utilizar.

Determinación de Puntos de Corte

El procedimiento para la determinación de las puntuaciones de corte, que sirvan para identificar los NL en la prueba, se basa en una adaptación del método *Bookmark*, teniendo en cuenta algunas variantes del método de Angoff, y consideraciones empíricas de los métodos utilizados por De la Orden (1998) y por Gaviria y Tourón (2000). Así, el método utilizado se sustenta sobre la actuación de un comité de expertos que determinan, a partir de los resultados obtenidos en los reactivos del Excale, cuáles son los elementos de la prueba característicos de cada nivel de logro. Los detalles del procedimiento utilizado se describen a continuación.

Comité 2

El Comité 2, para determinación de PC está compuesto por profesores en ejercicio, conocedores del comportamiento de los contenidos curriculares y de los alumnos del grado escolar correspondiente. El número de miembros del comité es de aproximadamente de cinco a ocho participantes.[15] En éste se integraron, como en el caso anterior, el director de la prueba como coordinador del mismo.

Forma de trabajo

El trabajo de este comité se organiza de la siguiente manera:

1. **Sesión de formación**. Desarrollada por el conductor del seminario—en sesión plenaria de los cuatro comités[16]— y por el coordinador de prueba —en sesión interna de cada comité— tiene como finalidad explicar a los miembros del comité el propósito general 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 que se familiaricen con el Excale correspondiente, los miembros del comité 2 responden a una prueba completa; con ello se pretende que tengan la oportunidad de conocer los contenidos de los reactivos, sus niveles de dificultad y los procesos intelectuales necesarios para responder a cada uno de ellos antes de juzgarlos.
3. **Sesiones de juicio**. Se establecen tres sesiones de juicio entre las que se aporta retroalimentación al comité acerca de sus niveles de acuerdo/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 que debe tenerse en cuenta es que se trata de identificar varias puntuaciones de corte a lo largo de un continuo. De esta forma, es previsible que se puedan dar dos tipos de discrepancias:

Figura 2. Ejemplo de un ítem del Cuaderno de Reactivos Ordenados (CRO): 3º de primaria, Matemáticas

No. de Reactivo: 42 Posición en la prueba: 15

Laura tiene $45.60, Carmen $55 y Miguel $27.75. ¿Cuánto dinero tienen entre los tres?

a) $128.35
b) $127.135
c) $73.90
d) $12 835

[15]Es deseable contar entre diez y veinte docentes por prueba, con el fin de conseguir suficiente representatividad sin arriesgarse a una excesiva dispersión, y para permitir una mayor estabilidad en los resultados de los análisis estadísticos. No obstante, la logística de este tipo de procesos obliga en ocasiones a tener números más reducidos de participantes. Téngase en cuenta que supone contar con personal que deja su puesto de trabajo habitual durante unos días y que proviene de diferentes estados del país.

[16]El modelo se aplicó por primera vez para cuatro asignaturas a la vez: Español y Matemáticas (6º de primaria y 3º de secundaria).

- 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 apartan significativamente del conjunto de estimaciones.[17] Esta estrategia es frecuente y persigue estimaciones más robustas y representativas.
- En las puntuaciones de corte. Aquí se pueden dar 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 dirige a los siguientes aspectos:

- Grado de congruencia entre los expertos para cada PC, mediante indicadores univariados para cada PC en cada sesión de juicio.[18]
- Número de reactivos que definen 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 recogen informaciones de los participantes acerca del desarrollo del mismo.

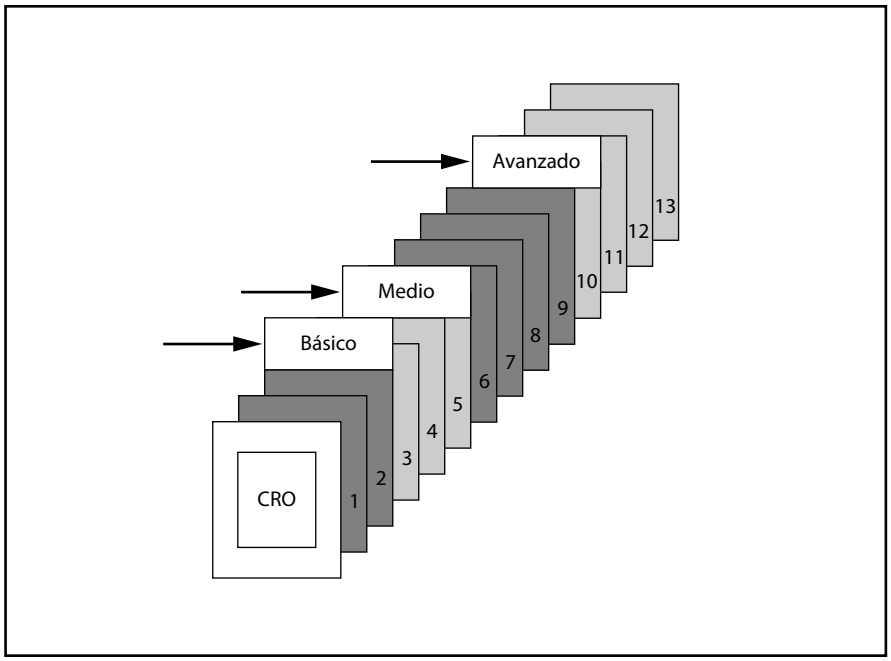
Formato de juicio

Para la emisión de juicios se requiere que cada miembro del comité trabaje sobre un cuadernillo de reactivos ordenados (CRO), el cual incluye los reactivos del Excale ordenados de menor a mayor dificultad (ver figura 3). Cada reactivo se presenta completo, identificando su nivel de dificultad.

La tarea que se plantea a los participantes es identificar cuáles reactivos pertenecen a cada una de las categorías de logro, comenzando por el reactivo más fácil y por la categoría de menor nivel de logro. Para ello, deben examinar cada reactivo y responder si un estudiante de la categoría que se está valorando en ese momento es capaz de responder correctamente al mismo.

Así, la pregunta que debe responder para cada reactivo es "¿un alumno del nivel θ puede responder correctamente este reactivo?". La respuesta que

Figura 3. Ilustración de CRO y reactivos marcadores



[17]Dado el reducido número de participantes en cada comité, en este caso no se utilizaron estrategias de identificación de jueces que aportan valoraciones extremas y por ello no se procedió a la eliminación de juicios como medida para lograr estimaciones robustas.

[18]En la validación del proceso se consideran aproximaciones univariadas, además de otras multivariadas para la validación del producto.

debe dar cada participante es SI/NO. No obstante, debe 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 serían, tomando como referencia una probabilidad de al menos el 67% de ellos (es decir, al menos, dos de cada tres alumnos).

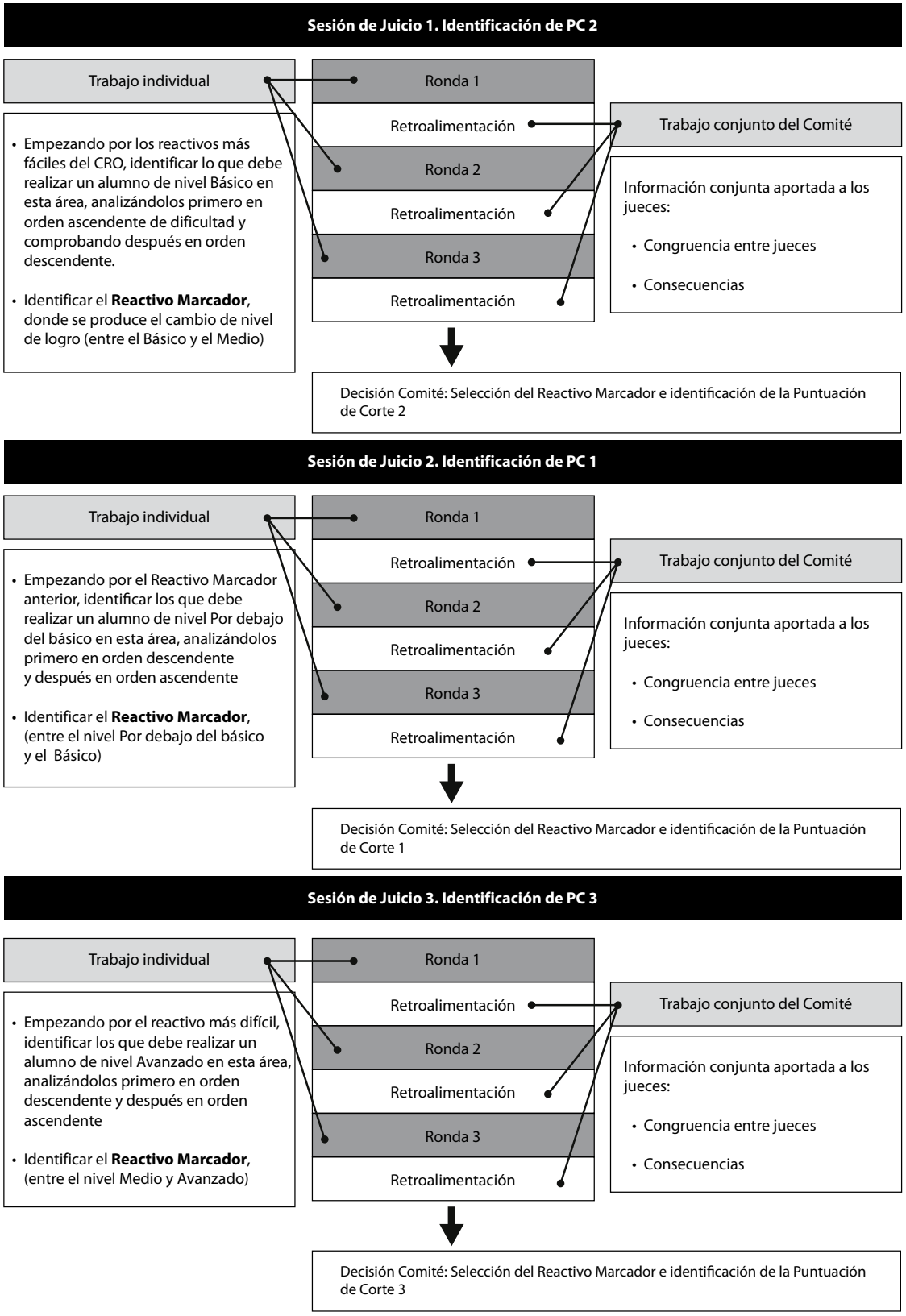
El cambio de un nivel de logro a otro se produce cuando se identifica un reactivo que un estudiante promedio del nivel actual no lo pueda responder. Ese reactivo actúa como marcador para identificar el PC. En caso de que se identifique un reactivo de cambio de nivel, pero posteriormente se dude acerca de si algunos reactivos de mayor dificultad podrían ser correctamente resueltos por estudiantes del nivel anterior (hecho poco posible, pero probable), cada participante debe revisar su reactivo marcador hasta asegurarse que esté situado en el nivel más representativo. No obstante, cada participante debe identificar tantos reactivos marcadores como PC haya entre niveles. Como la categorización que utiliza el INEE es de cuatro niveles, para las pruebas Excale se tienen que identificar tres marcadores (ver figura 3 y tabla VIII).

Una esquematización del proceso de juicio, en la que se indican tanto las tareas a realizar por los especialistas como el sistema de rondas de juicio, se presenta en la figura 4.

Tabla VIII. Esquematización de niveles y puntuaciones de corte θ_1

Niveles	Puntuaciones de corte
Por debajo del básico	→ θ_1
Básico	→ θ_2
Medio	→ θ_3
Avanzado	

Figura 4 . Sesiones de juicio del Comité 2.



Criterios de decisión e identificación PC

Teniendo en cuenta la experiencia de procesos similares que se reporta en la literatura especializada, usualmente con tres sesiones de juicio normalmente se llegan a determinar los puntos de corte 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 respecto al proceso y a los niveles resultantes. Por este motivo, se toma como esquema de trabajo realizar tres rondas de juicio (ver figura 4).

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 PC aquélla que resulte del máximo nivel de congruencia en la opinión de los docentes y, por lo tanto, se pueda entender como un estimador robusto del consenso intersubjetivo.

Como criterios de calidad para considerar concluido el proceso se debe tener en cuenta lo siguiente: 1) los niveles de congruencia entre los participantes para cada PC y 2) 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).

Protocolo de actuación

Como en el caso del Comité 1, el trabajo del Comité 2 se organiza en función de protocolos de actuación. Presentamos aquí los relativos a: 1) formación del comité, 2) emisión de juicios, que incluye el formato de juicio, así como los modos en que se aporta información para la retroalimentación al comité y 3) criterios de selección de PC.

Recuadro 4

Protocolo para la formación del Comité 2

La formación que se da al Comité 2 para realizar su tarea, contempla los siguientes contenidos:

1. Explicar la finalidad y objetivos de la determinación de los NL.
2. Mostrar las definiciones realizadas por otras instituciones. En cada caso, con ejemplos de la materia sobre la que trabajaba cada comité.
3. Presentar 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. Presentar el formato de juicio, indicando claramente el procedimiento a seguir, especificando que se realizarían diversas rondas de juicio para la determinación de cada PC, el tipo de retroalimentación que se ofrece y los criterios de convergencia y control que se debe utilizar.
5. Mostrar un ejemplo simulado.
6. Presentar del CRO[20] y explicar la información que corresponde a un reactivo.
7. Concluida la sesión de formación, administrar la prueba sobre la que se van a identificar los PC. La finalidad es que puedan tomar contacto real con la prueba. A los miembros del comité se les presenta una prueba completa, y se les solicita que la respondan. Posteriormente se les aporta la clave de respuestas con el fin de que comprueban si han fallado algún reactivo; la finalidad de esta actividad es que tengan la oportunidad de comprobar la calidad, claridad y niveles de dificultad de los reactivos que iban a juzgar posteriormente.

Recuadro 5

Protocolo para el formato de juicio

1. A cada miembro del comité se le facilita un CRO.
2. Cada miembro emite 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, puede aceptarse un debate acerca de sus implicaciones.
4. Cada sesión de juicio se centra sobre la identificación de un único PC, con la secuencia que posteriormente describiremos. Pueden realizarse hasta tres rondas de juicio, con el fin de facilitar la convergencia entre los jueces.[1]
5. Al finalizar cada ronda, se introducen los datos y se realiza un breve análisis en el que se comprueba: a) la convergencia entre jueces y b) las consecuencias de la aplicación del PC para describir los resultados de la evaluación.

   Dicha información se ofrece como *feedback* (cuyo protocolo también describiremos posteriormente) a los participantes que pueden revisar su juicio anterior.
6. La tarea que se plantea a los participantes es: comenzando por el reactivo más fácil y por la categoría de menor nivel de logro, se trata de que identifiquen qué reactivos pertenecen a cada una de las categorías. Para ello, deben examinar cada reactivo y responder si un sujeto de la categoría que se está valorando en ese momento es capaz de responder correctamente al reactivo.
7. La pregunta que se debe responder para cada reactivo es: "¿un sujeto del nivel θ puede responder correctamente este reactivo?". La respuesta que debe dar cada participante es SI/NO, como hemos indicado anteriormente.
8. En términos generales, la identificación del reactivo en el que se produce el primer "NO" indica el cambio de NL: es justamente el reactivo marcador, que se produce cuando el juez identifica un reactivo que estima poco probable que lo pueda responder un sujeto del nivel actual. En otras palabras, ese reactivo actúa como marcador para identificar el PC, que debe representar el nivel mínimo que deben mostrar los alumnos para ser considerados dentro de NL determinado.
9. En caso de que se identifique un reactivo como marcador, pero posteriormente se dude acerca de si algunos reactivos de mayor dificultad pueden ser correctamente resueltos por sujetos del nivel anterior —o viceversa— (hecho poco posible, pero probable), cada participante debe revisar 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 y sean señalados como *incidencias en la determinación de estándares*.

   Estos casos pueden dar origen a recomendaciones como consecuencia de la evaluación, ya que podrían constituir interpretaciones complementarias de carácter cualitativo a la información meramente cuantitativa.
10. Asimismo, se tiene en cuenta que pueden darse este tipo de incidencias por diversos motivos, entre ellos podemos señalar los más frecuentes:
    - Un contenido teóricamente fácil puede estar medido por un reactivo mal diseñado, de forma que ello lo convierta en más difícil.
    - Puede resultar que dicho contenido no se imparta habitualmente en las clases, aunque esté presente en el currículo.
    - 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.

Recuadro 5 (continuación)

11. Dado que se trata de identificar tres puntuaciones de corte, y considerando que cada NL debe responder a competencias claramente delimitadas, la secuencia de identificación de puntuaciones se altera (desde la mínima a la máxima), en el siguiente orden:

$$\boxed{\theta_2, \theta_1 \text{ y } \theta_3}$$

donde: θ_2 = PC entre el nivel *Básico* y el *Medio*
θ_1 = PC entre el nivel *Por debajo del básico* y el *Básico*
θ_3 = PC entre los niveles *Medio* y *Avanzado*

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

La tarea para cada PC es:

- θ_2 se identifican desde los reactivos más fáciles, en orden ascendente (posteriormente se comprueban en sentido inverso). La tarea que deben realizar los participantes es identificar todos los reactivos que debe realizar un alumno cuyo nivel es *Básico*.
- 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*.
- θ_3 es la PC 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*.

12. Una vez obtenida la convergencia en torno a un PC se da por concluida la sesión, y se da paso, según el programa de actividades, a la sesión correspondiente al siguiente PC, hasta concluir el proceso.
13. Cada participante debe identificar tantos reactivos marcadores como PC entre niveles. De modo que, como la categorización es de cuatro niveles, deben identificase tres PC, es decir, tres marcadores, tal como indicamos anteriormente y se muestra en el esquema de la figura 4.

Recuadro 6

Protocolo para la retroalimentación

1. La información de retroalimentación tiene por objeto ayudar a orientar el acuerdo interjueces. 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 produce la información reúne los siguientes elementos:
   a) Congruencia entre jueces. Se indicó:
      - El rango de puntajes que se emitieron (por ejemplo, entre 12 y 15).
      - La variabilidad de los juicios expresada a partir de la desviación estándar.
      - El nivel de acuerdo que hay 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 indicó:
      - El porcentaje de sujetos que quedan por encima y por debajo de cada puntaje y del posible PC, estimado como la media de las aportadas.
      - Distancia en puntajes y en porcentaje de sujetos desde el PC y la media.
      - 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 IX y en las figuras 5 y 6 se ejemplifica el tipo de información de retroalimentación que se debe aportar.

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

Asignatura: Español **Nivel:** Básico

Asistencia	Juez	Reactivo marcador	Puntuación de corte
S	1	12	518.9
S	2	14	535.0
S	3	8	498.7
S	4	13	533.7
S	5	9	506.3

Ronda: 1

Indicador	Reactivo marcador	Puntuación de corte
PC1	12	518.9
Mínimo	8	498.7
Máximo	14	535.0
D.E.		16.2

Figura 5. Distribución de los jueces y reactivos (para PC1 sugerida)

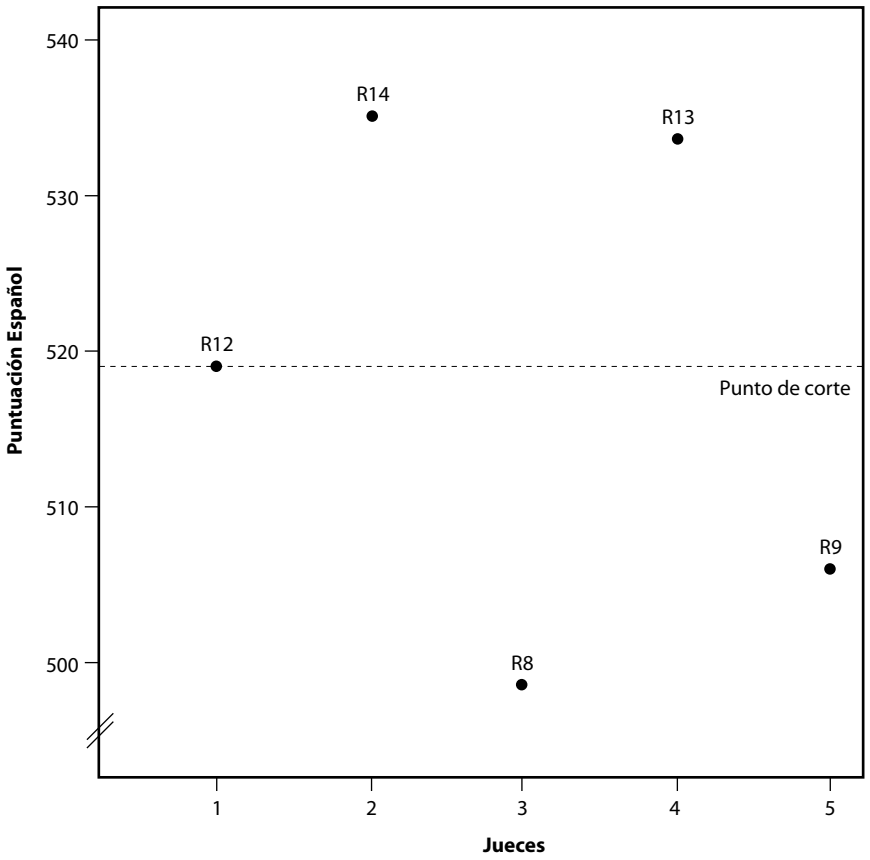
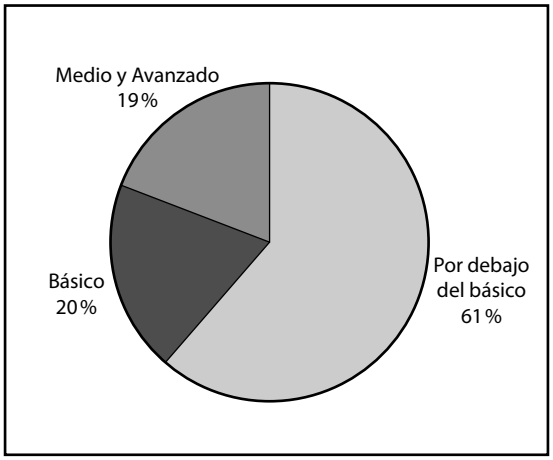
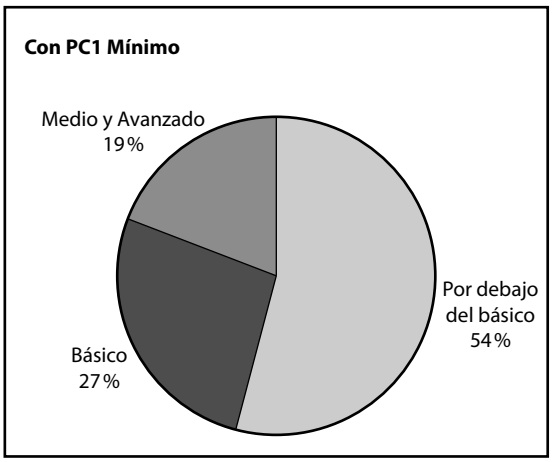
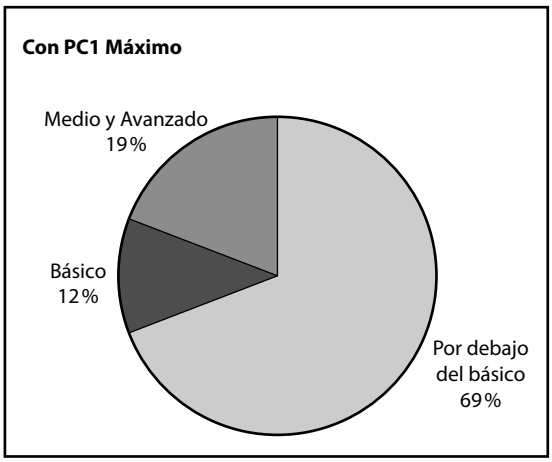


Figura 6. Distribución de estudiantes para el PC1 sugerido







Recuadro 7

Protocolo para la selección de los PC

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 tienen en cuenta:
   - Los niveles de congruencia entre los participantes para cada PC. Para este cometido, se toma como referencia la distancia entre las puntuaciones de corte y la desviación de los juicios emitidos.
   - La valoración por parte de los participantes acerca de la representatividad de los NL obtenidos en cuanto al porcentaje de sujetos identificados en cada nivel.
   - En cualquier caso, el ajuste final de cada puntuación se realiza de forma que se mantenga una distancia entre PC suficiente. Este aspecto, así como otros indicadores de calidad, se muestran en el apartado de validación.

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

Informes a emitir	Quién desarrolla la tarea		
	Equipo externo a los comités	Comité 1 (NL)	Comité meta evaluador
Informe de evaluación del proceso	Dirige, recoge y elabora información	Debate el informe y emite sugerencias	Revisa y analiza: • Documentación técnica • Informe • Metodología seguida en el proceso
Informe de validación del proceso			Valida o refuta el procedimiento

Validación de los Niveles de Logro

La validación de los Niveles de Logro se realiza considerando diversos tipos de evidencias, tanto del proceso mismo para la determinación de los estándares, como de la definición de los propios Niveles de Logro. En este sentido, la validación de los estándares para interpretar los resultados de los Excale también forma parte de la metodología utilizada. La validez de los NL como sistema de interpretación de los resultados de las pruebas depende de múltiples factores, y el INEE, como responsable último de la evaluación, es quien debe priorizar cuáles son las evidencias fundamentales que se deben recoger para utilizar de manera óptima la información de sus evaluaciones de aprendizaje. No obstante, la evidencia básica que se incluye en el modelo es la basada en *la evaluación del proceso de determinación de los NL*.

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 NL 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 del modelo utilizado, revisamos los diversos componentes del mismo.

Objeto y finalidad de la evaluación

Se trata de evaluar si el proceso de determinación de los NL se ha desarrollado de forma adecuada. La finalidad es doble: 1) formativa, de manera que durante el proceso se trataron de corregir problemas detectados durante el desarrollo del mismo y b) sumativa, como rendición de cuentas (y evidencia de validez) acerca de la representatividad y calidad de los NL identificados como sistema de interpretación de puntuaciones.

Así, el objetivo principal de esta parte del proceso es recabar información acerca de la calidad del proceso para la determinación de Niveles de Logro y del establecimiento de puntuaciones de corte.

Como objetivos específicos, tenemos los siguientes:

- Documentar y analizar el proceso desarrollado, garantizando la ausencia de sesgos en el mismo y valorando su adecuación metodológica.
- Analizar la calidad de los Niveles de Logro establecidos y de los puntos de corte seleccionados.
- Analizar la influencia de variables extrañas en el proceso que haya podido deteriorar los resultados.

Evaluación e informes

Para cumplir con estos objetivos, el proceso de evaluación contempla dos etapas: como se muestra en la tabla X.

1. Elaboración de un informe de evaluación dirigido y desarrollado por un equipo externo de especialistas, 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 participa en las sesiones a través de un observador externo que funge como asistente del coordinador de prueba, si bien su rol se centra exclusivamente en las tareas de evaluación.
2. Comité meta evaluador (CME),[19] encargado de revisar la información derivada de la evaluación,

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

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 NL como en la evaluación. Como resultado de su actuación, emite un informe final de validación del proceso, en el que aporta evidencias necesarias que apoyan o refutan la calidad del procedimiento desarrollado. Este rol recae, en última instancia, en las autoridades del INEE.

3. 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.

Las fuentes de información que se consideran son las siguientes:

1. Los participantes, en al menos tres aspectos: a) el análisis de sus respuestas de identificación de PC, b) el conocimiento y comprensión de los métodos y procedimientos a utilizar y c) sus opiniones acerca del proceso.
2. El coordinador de prueba: su valoración acerca del proceso.
3. El observador externo: su valoración acerca del proceso.
4. El comité meta evaluador: su valoración metodológica (validación del informe de evaluación).

Las variables e indicadores que se tuvieron en cuenta en este proceso, así como las fuentes de información de las que se extrajeron aparecen en la tabla XI.

Tabla XI. Síntesis de indicadores y fuentes de información

Tipo	Indicadores	Fuentes de información				
		Valoraciones de NL	Opiniones de participantes	Coordinador de prueba	Observador externo	Otras fuentes
Entrada	Características profesionales de los participantes[1]		X	X		
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 confiabilidad 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 NL identificados de una a otra sesión de juicio	X				
	Satisfacción con el proceso de formación	X	X	X	X	
Resultado	Satisfacción con los procedimientos utilizados	X	X	X	X	
	Satisfacción con el funcionamiento global del comité	X	X	X	X	
	Congruencia en la identificación de puntuaciones de corte (en cada sesión de juicio) Perspectivas univariada y multivariada	X				
	Confiabilidad asociada a las puntuaciones de corte en cada nivel	X				
	Distribución porcentual de los sujetos en los NL	X				
	Satisfacción con la adecuación de los NL determinados	X	X	X	X	
Contexto	Comparación del funcionamiento de los diversos comités de las diferentes materias					X
	Análisis lógico de los NL identificados para cada materia con los utilizados en otro proyectos evaluativos comparables					X

Respecto a los *momentos de recopilación de información*, las variables e indicadores *de entrada* se recaban previamente, o al inicio de las sesiones de juicio. Hay que distinguir entre indicadores y variables *de proceso*, así el indicador relativo a la comprensión de la tarea y procedimientos se recoge 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 extraen tres medidas.

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 y tiene valor contextual.

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

1. Variables de entrada, recogidas mediante cuestionario dirigido a los participantes en los comités
2. Valoración del conocimiento y comprensión de tareas y métodos, mediante prueba estandarizada dirigida a los miembros de los comités
3. Tasas de cambio en indicadores y niveles de congruencia, confiabilidad, características de las distribuciones, tanto las variables de proceso como las de producto: datos de carácter estadístico, tanto a nivel univariado como multivariado
4. 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 comprueba la adecuación de la información y las valoraciones contenidas en él, de forma que emite un juicio valorativo global que atiende a tres aspectos:

1. La calidad del proceso desarrollado para la determinación de Niveles de Logro.
2. La adecuación metodológica y de las decisiones tomadas.
3. La calidad de los NL identificados.

Para esta evaluación, el CME dispone de toda la documentación disponible acerca de los trabajos realizados por el CSI y CNL, así como del informe de evaluación, y puede recabar (si lo estima oportuno) la información complementaria que precise, tanto documental, como de audiencias con participantes.

Metodología

El trabajo con comités de especialistas es un área emergente, que progresivamente va siendo cada vez más utilizada en diversos ámbitos de la evaluación. Esto es especialmente cierto para el campo del desarrollo de pruebas educativas de gran escala, y su uso es muy conveniente en tareas relacionadas con el análisis y especificación de dominios educativos (como universos de medida), la construcción y revisión de reactivos, así como en la determinación de NL y PC. Al tratarse del trabajo con grupos pequeños de especialistas, en ellos confluyen diversas aproximaciones metodológicas, tanto de carácter cuantitativo como cualitativo, y tanto para la recopilación de información como para su análisis y síntesis. En nuestro caso, se combinan diversas estrategias como elementos de:

- Recopilación de información en este modelo, a través de cuestionarios semiestructurados y observación no sistemática (diarios de observadores y registros observacionales de coordinadores de comités).
- Análisis y síntesis de información estadística y reseña de observaciones abiertas.
- Validación de la interpretación y conclusiones, donde se contrastan diversas fuentes de información.

En términos generales, podemos identificar el estudio de validación en el marco de los estudios observacionales basados en metodologías complementarias.

Indicadores e instrumentos

Los indicadores considerados en este proceso se recogen en la tabla XII, donde se sintetizan los diferentes aspectos de la validación del proceso. En este apartado reseñamos los instrumentos asentados en la tabla XII y realizamos una breve descripción de los mismos, basándonos en las dimensiones de síntesis para la interpretación de la información recabada a través de cada uno de ellos.

Como puede observarse, se han desarrollado tres cuestionarios y diversas hojas de registro para observaciones. Pasamos a describir su contenido brevemente.

Cuestionario 1: Consta de 17 reactivos y está dirigido a recoger las opiniones de los participantes en el Comité 1; se administra al final del proceso al concluir la segunda sesión de trabajo. Las dimensiones de síntesis de información consideradas son: infraestructura de trabajo, adecuación de la formación inicial, actuación del coordinador, calidad de los NL, organización general del seminario, utilidad de los debates de grupo y valoración de las propuestas del Comité 2.

Tabla XII. Síntesis de instrumentos para evaluar el proceso de NL y PC

Instrumento	Fuente de información		Momento de la aplicación
Cuestionario 1	Comité 1		Final del proceso
Cuestionario 2.1	Comité 2		Final de la formación
Cuestionario 2.2	Comité 2		Final del proceso
Hoja de registro de valoraciones	Comité 2	Puntuación de corte 1	Durante el proceso
		Puntuación de corte 2	
		Puntuación de corte 3	
Hoja de registro de observaciones	Coordinador Comité 1		Durante el proceso
Hoja de registro de observaciones	Coordinador Comité 2		Durante el proceso
Observación no sistemática (diario de sesiones)	Observadores/evaluadores		Durante el proceso

Cuestionario 2.1: Consta de 14 reactivos y está dirigido a recoger las opiniones de los participantes en el Comité 2 y a valorar la comprensión de la tarea tras la formación inicial. Se administra al concluir el proceso de formación. Su finalidad es formativa; se trata de recabar información para asegurar que el proceso se vaya desarrollando adecuadamente y, en su caso, introducir las correcciones oportunas. Las dimensiones de síntesis de información consideradas son: 1) satisfacción general acerca del proceso de formación, 2) conocimiento y comprensión de la tarea a realizar y 3) valoración de la definición de los NL planteados por el Comité 1.

Cuestionario 2.2: Consta de 14 reactivos y está dirigido a recoger las opiniones de los participantes en el Comité 2. Se administra al finalizar todo el proceso de identificación de las puntuaciones de corte. Su finalidad es recabar las opiniones de los participantes acerca de la adecuación general del proceso seguido y de los resultados obtenidos. Las dimensiones de síntesis[20] de información consideradas son: 1) infraestructura de trabajo, 2) adecuación de la formación inicial, 3) actuación del coordinador, 4) organización general del seminario, 5) utilidad de los debates de grupo, 6) identificación de los PC, 7) información de retroalimentación, 8) la seguridad acerca de cómo han seguido los participantes las instrucciones y 9) claridad y utilidad de los descriptores aportados por el Comité 1.

20Aunque las identificamos aquí (y en lo sucesivo) como dimensiones, corresponden a dimensiones de contenido no identificadas mediante ningún análisis de reducción de datos. Algunas de ellas están representadas por un solo reactivo.

Hojas de registro de valoraciones: (para puntuaciones de corte): Están dirigidas a recoger los juicios de cada participante para identificar las puntuaciones de corte, y recabar información acerca de este proceso (dificultades en la identificación, reactivos desubicados en cuanto al nivel, etc.). Se cumplimentan al finalizar cada sesión de juicio. Son tres, una para cada PC.

Hojas de registro de observaciones de coordinadores: Presentan diversos formatos y están orientadas a recoger las apreciaciones de los coordinadores acerca del trabajo de los dos comités. Se incluyen hojas de registro para ambos comités.

Análisis de resultados

Como informantes, se incluyen a todos aquellos que han tenido un rol específico en el funcionamiento interno de los comités y que han desarrollado una tarea concreta en relación con el objeto de trabajo del comité o de la validación del proceso. Se pueden distinguir cuatro tipos de participantes: 1) participantes en los comités tipo 1 y 2, 2) coordinadores de los comités, 3) apoyo técnico y 4) observadores y evaluadores.

Respecto a los resultados, podemos sintetizarlos en las siguientes aproximaciones:

1. Análisis de cuestionarios de opinión (1, 2.1 y 2.2). Se realizan análisis descriptivos de los reactivos de cada uno de ellos, así como de las dimensiones de contenido en que pueden sintetizarse. Este tratamiento se aplica sobre los totales de los comités, así como se segrega el análisis para

cada uno por los comités de cada materia, se realizan comparaciones entre los niveles medios. Dado el escaso número de observaciones que suele darse en este tipo de procesos, se utilizan dos aproximaciones con objeto de asegurar la aplicacibilidad de los estadísticos utilizados y, en su caso, valorar la concurrencia de resultados: *t de Student* —incluyendo un contraste previo de homogeneidad de varianzas, mediante la prueba de Levène—, y la prueba U de Mann-Whitney (como opción no paramétrica más ajustada al tipo de datos disponibles). Estas opciones de contraste deben entenderse como elementos de orientación para el establecimiento de conclusiones y ser tomadas con la cautela necesaria en consonancia con los datos disponibles. Adicionalmente, se sintetizan las respuestas abiertas aportadas por los participantes en los comités.

2. Observaciones de coordinadores y observadores. Se realizan las síntesis de observaciones recabadas en registros y diarios de sesiones. Dichas síntesis se realizan en contraste con los implicados en cada caso. Así, se consensúa con los coordinadores sus informaciones, y se realiza la síntesis de observadores por acuerdo entre el equipo de observadores/evaluadores que actúa en el proceso. De manera adicional, se realizan dos sesiones conjuntas de puesta en común y síntesis de información con el conjunto de coordinadores y observadores implicados.
3. Identificación de puntuaciones de corte: Se utilizan diversos indicadores de convergencia de los juicios emitidos cuya finalidad es valorar la calidad de las puntuaciones de corte. La escala —establecida mediante la Teoría de Respuesta al Ítem (IRT, en inglés)— tiene una media de 500 y una desviación de 100. Como indicadores tenemos:
   a. Precisión del juicio: En cada ronda de juicio se valora la desviación de los juicios respecto a la mediana como indicador base para valorar la distancia de los juicios emitidos al PC seleccionado. Este indicador tiene como referente para orientar el criterio el valor mismo de la desviación de la escala (100 puntos), de forma que puede entenderse que una $\sigma = 0$ indica convergencia total de juicios. Aunque no hay referencias precisas del indicador.
   b. Razón de acuerdo (RA) entre jueces al determinar un PC. Se estima como el porcentaje de jueces que coinciden en la identificación de una puntuación. En este Modelo, hemos tomado diferentes intervalos para valorar la coincidencia de juicios: $\sigma \pm 2.5\%$, $\sigma \pm 5\%$, $\sigma \pm 7.5\%$, $\sigma \pm 10\%$, $\sigma \pm 12.5\%$, $\sigma \pm 15\%$,..., $\sigma \pm 25\%$. En cada caso se contabilizan los jueces que emiten valoraciones en cada tramo. Teniendo en cuenta los resultados que se evidencian en la literatura especializada, se considera un buen nivel de convergencia una $RA \geq 70\%$, al menos en un intervalo de $\sigma \pm 10\%$.
   c. Sesgos de valoraciones (SV). Se estiman las distancias medias que se producen por encima y por debajo de la PC a fin de analizar las tendencias de valoración que se han producido al determinar la PC, con el fin de valorar su robustez final. Entre los PC no convergentes se considera una estimación sin sesgo aquella cuyas distancias entre juicios superiores e inferiores son equidistantes; por el contrario, una estimación sesgada es aquella donde las puntuaciones extremas no son equidistantes
4. Por otra parte, se tiene en cuenta la representatividad de los Niveles de Logro obtenidos en relación con el porcentaje de estudiantes que se ubican en cada nivel, así como las opiniones de los participantes sobre el proceso de establecimiento de cada PC. Ambas informaciones se sintetizan mediante análisis estadísticos descriptivos.

En las tablas XIII, XIV y XV se presentan algunos ejemplos de resultados obtenidos para el Excale de sexto de primaria.

Tabla XIII. Ejemplo de síntesis de resultados del cuestionario 1

Dimensiones de valoración	Muy negativo	Bastante negativo	Negativo	Positivo	Muy positivo
Actuación del coordinador del seminario					M1, M2 E1, E2
Infraestructuras para el trabajo					M1, M2 E1, E2
Organización general		M1	E2	M2, E1	
Debates de grupo					M1, M2 E1, E2
Niveles de logro			E2	M1, M2 E1	
Formación inicial				E1, E2	M1, M2
Valoración de las propuestas del Comité 2			M1, M2 E1, E2		

Donde: M1 = Comité Excale-06 / Matemáticas, M2 = Comité Excale-09 / Matemáticas, E1 = Comité Excale-06 / Español, E2 = Comité Excale-06 / Español. Las celdas sombreadas indican las medias de las puntuaciones.

Tabla XIV. Ejemplo de síntesis de resultados del cuestionario 2.1

Dimensiones de valoración	Muy negativo	Bastante negativo	Negativo	Positivo	Muy positivo
Satisfacción con la formación inicial				M1, E1 E2	M2
Conocimiento y comprensión de la tarea				M1, E1	M2 E2
Utilidad de descriptores de los niveles de logro del Comité 1			E1	M1	M2, E2

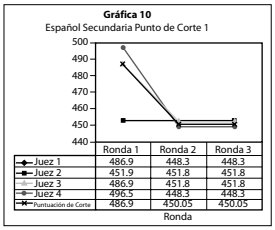
Donde: M1 = Comité Excale-06 / Matemáticas, M2 = Comité Excale-09 / Matemáticas, E1 = Comité Excale-06 / Español, E2 = Comité Excale-06 / Español. Las celdas sombreadas indican las medias de las puntuaciones.

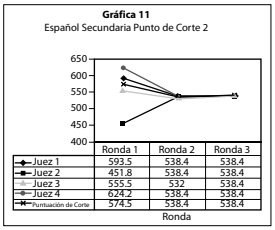
Tabla XV. Ejemplo de síntesis de resultados del cuestionario 2.2

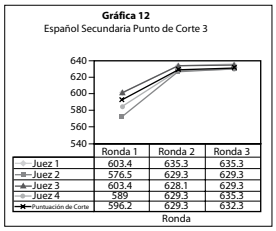
Dimensiones de valoración	Muy negativo	Bastante negativo	Negativo	Positivo	Muy positivo
Actuación del coordinador del seminario					M1, M2 E1, E2
Infraestructuras para el trabajo					M1, M2 E1, E2
Organización general				E2	M1, M2 E1
Debates de grupo					M1, M2 E1, E2
Información de retroalimentación				E1, E2	M1, M2
Formación inicial				E2	M1, M2 E1
Identificación de puntuaciones de corte			E2	M1, M2 E1	
Seguridad en seguir instrucciones				M1, M2	E1, E2
Utilidad de descriptores de los NL del Comité 1			E1	M1, M2 E2	

Donde: M1 = Comité Excale-06 / Matemáticas, M2 = Comité Excale-09 / Matemáticas, E1 = Comité Excale-06 / Español, E2 = Comité Excale-06 / Español. Las celdas sombreadas indican las medias de las puntuaciones.

Figura 7. Ejemplo de resultados de indicadores de calidad del proceso de identificación de PC: Excale -09 / Español



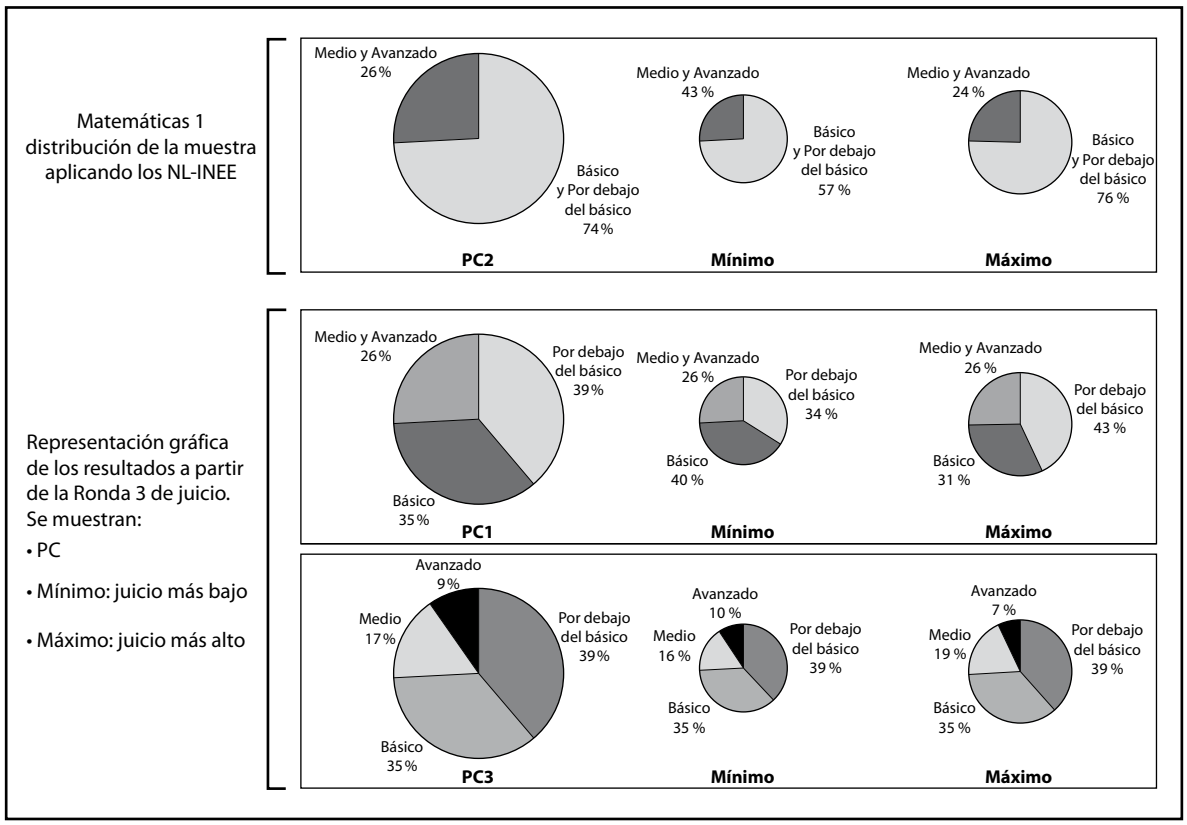




	Español 2. Indicadores de calidad de la identificación de las PCs.		
	Puntuación de Corte		
Indicadores	PC1	PC2	PC3
Convergencia	No	Sí	No
σ de juicios	2.00	0.00	3.50
Rango de juicios	3.50	0.00	6.00
Porcentaje jueces en *PC ± 5% σ*	100	100	100
Observaciones			

Nivel de logro	θ	Porcentaje de alumnos
Por debajo del básico	450.05	33
Básico	538.40	35
Medio	632.30	25
Avanzado		7

Figura 8. Ejemplo de distribución de PC identificados en tres momentos: Excale-06 / Matemáticas



CONCLUSIONES Y RECOMENDACIONES DE MEJORA DEL MODELO

Un elemento en el que se sustenta el trabajo de construcción de pruebas basado en juicios de especialistas es la calidad profesional y académica de cada uno de ellos (Hambleton, 1998, 2001; Cizek, 2001; Cizek, Bunch y Koons, 2004). En el caso del INEE, los especialistas que han conformado los distintos comités han sido seleccionados tomado en cuenta las siguientes características: 1) que representaran diversas condiciones escolares de importancia para el Sistema Educativo Mexicano, atendiendo a variables tales como el tipo de escuela, la modalidad educativa y el tipo de sostenimiento, 2) la antigüedad en el trabajo y su prestigio profesional en temas de docencia, innovación educativa, evaluación de la educación y 3) su participación previa en los diversos trabajos que realiza el INEE (diseño de la prueba, construcción de reactivos, validación y ausencia de sesgo de reactivos, etcétera).

En cuanto a las variables de proceso pueden identificarse dos grandes grupos: a) las relativas a las opiniones de los participantes en el proceso, y b) las que corresponden con la identificación de las puntuaciones de corte. Con este mismo esquema las comentamos a continuación.

COMITÉ 1: NIVELES DE LOGRO

En general, en los procesos de identificación de NL en que se ha aplicado este modelo, los elementos fuertes del mismo han sido la actuación del coordinador, las infraestructuras de trabajo y la utilidad de los debates de grupo, la formación inicial, y la valoración que realizan acerca de los estándares producidos. Los elementos débiles del modelo, según los participantes, son: la valoración de las propuestas del Comité 2, en opinión del comité 1 y la organización general de los comités.

Respecto a las apreciaciones que se derivan de las observaciones de coordinadores y evaluadores, se puede afirmar que las conclusiones son concurrentes, de forma que se identifican los mismos elementos fuertes y débiles. En este último caso, y considerando las respuestas dadas por los participantes, así como las observaciones de coordinado-

res y evaluadores, el factor más negativo de la organización de la experiencia fue la falta de tiempo, en especial en la primera sesión de estos comités, de forma que en alguno de ellos se tuvo que ampliar el horario de trabajo.

Comité 2: Puntos de Corte

En estos comités se recabó información en dos momentos: tras la formación inicial y al finalizar el proceso. Respecto al primer momento de valoración, en general los elementos fuertes del modelo son: la formación inicial del comité con la propuesta inicial de categorías y etiquetas que propone el INEE para el desarrollo de los NL. Estas valoraciones son concurrentes con las observaciones de coordinadores y evaluadores, si bien se pueden realizar algunos matices al respecto derivadas de la observación del trabajo de los comités.

Así, se aprecia que es especialmente importante el concepto de reactivo marcador (donde suelen manifestarse dudas entre los participantes acerca de si es el primero o el último de un nivel), así como la comprensión de la información de retroalimentación y cómo actuar en consecuencia con ella (se observa que hay una dificultad clara para relacionar las decisiones consecuentes a esta información, por la falta de comprensión de las relaciones entre la distribución de sujetos que se aprecia en los gráficos de sectores y la dirección que debe tomarse en la búsqueda del reactivo marcador en la siguiente ronda). No obstante, estas dificultades se pueden subsanar durante el proceso, de forma que los coordinadores de comités deben estar especialmente atentos a este tipo de dificultades.

Respecto a la valoración realizada al finalizar el proceso, los elementos fuertes del proceso son: la actuación de los coordinadores de comités, la infraestructura para el trabajo, la utilidad de los debates de grupo, la formación inicial, la organización general y respecto a la seguridad relativa al modo en que se siguen las instrucciones de trabajo. Asimismo, se pone de manifiesto la utilidad de la información de retroalimentación y la identificación de las puntuaciones de corte y la utilidad de los descriptores aportados por los comités NL. En términos generales, y respecto a este segundo momento de valoración, las posiciones de los participantes son concurrentes por las observaciones de coordinadores y evaluadores.

Por otra parte, y en relación con el proceso específico de identificación de las puntuaciones de corte, en todos los comités se pusieron de manifiesto las ventajas del trabajo a partir de protocolos de actuación, dado que homogeneizan en lo sustancial el sistema de trabajo, de forma que las diversas pruebas de los Excale están sujetas al mismo tipo de procedimiento. El número de rondas de juicio fue habitualmente el previsto (de tres en todos los comités), por lo que el modelo se puede considerar bien ajustado en relación a este aspecto. Un elemento a resaltar es que las modificaciones de juicio de una a otra ronda se producen especialmente por la representatividad de la distribución de los alumnos al aplicar los estándares producidos.

Otro elemento importante para la calidad de las pruebas es que al utilizar este modelo, se pueden identificar los reactivos desubicados respecto a los NL planteados.

Por otro lado, en la determinación de las puntuaciones de corte, este modelo ha permitido un alto nivel de congruencia. La distribución porcentual de los alumnos en cada nivel de logro, permite que los participantes analicen si ésta es representativa de la distribución de alumnos que se suele observar en el salón de clases. Y, en consecuencia, diferenciar si los niveles de dificultad observados en algunas pruebas se deben a la población de referencia o a las características de las pruebas.

Acerca de los elementos a mejorar en el proceso de trabajo diseñado para el modelo

La estructura global de trabajo en las diversas aplicaciones del modelo se considera positiva. Sin embargo, a partir de la información recabada durante el proceso de validación se pueden extraer las siguientes recomendaciones de mejora.

Respecto a los comités de NL, la estructura temporal inicial es corta; se requiere una primera sesión de dos jornadas de trabajo, de forma que en la primera pueda destinarse además del tiempo de formación, un tiempo adicional para el conocimiento y profundización en los materiales de trabajo, y así pueda disponerse de una jornada completa para la redacción inicial de descriptores. El momento 2 de trabajo está bien ajustado.

1. Respecto a los comités de puntos de corte, es conveniente aprovechar las tres jornadas de trabajo, de forma que la primera se pueda destinar por completo a la formación y práctica de sus miembros. De este modo, al protocolo de formación inicial es necesario añadir ejercicios prácticos de uso de la información de retroalimentación y su vinculación con decisiones posteriores.

2. En todos los casos, es conveniente que los comités de una misma materia, pero de distintos grados, puedan disponer de una sesión final conjunta, de manera que pueda asegurarse la continuidad de niveles entre materias, tanto entre el Comité 1, como en el Comité 2.

Consecuencias del modelo para mejorar el diseño de las pruebas

La experiencia desarrollada ha puesto de manifiesto diversos elementos que es conveniente tener en cuenta en el diseño y desarrollo de pruebas de estas características. En síntesis:

1. La relación entre el análisis reticular (análisis gráfico del currículo, donde se establecen relaciones de servicio entre los distintos temas) y el desarrollo de Niveles de Logro debe ser iterativo, no secuencial como se ha venido haciendo en el INEE; de forma que el establecimiento de estándares pueda realimentar el establecimiento de la retícula y viceversa.
2. En consecuencia, el proceso de desarrollo de Niveles de Logro debe asociarse a las primeras etapas del desarrollo de las pruebas, durante el proceso de definición del dominio educativo como universo de medida. Ello conllevaría, sin duda, mejoras para la identificación de las claves de discriminación entre niveles y, con ello, podría incidir en la escritura de reactivos, su selección y revisión.
3. Es conveniente, asimismo, vincular estos trabajos a procesos de pilotaje de las pruebas, para que puedan probarse los puntos de corte representativos de dichos niveles. Ello posibilitaría añadir alguna experiencia de validación basada en métodos de determinación de NL que toman como referencia a los estudiantes y no a los reactivos.

En definitiva se trata de acercar el diseño y desarrollo de las pruebas alineadas al currículo con las pruebas referidas a estándares, lo cual entendemos que es compatible y no perjudica la representatividad curricular de las mismas.

Bibliografía

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

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*: Erlbaum, Mahwah, NJ, pp. 373-386.

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*. Erlbaum, Mahwah NJ, pp. 445-476.

Carson, J.D. (2001): Legal issues in standard setting for licednsure. In G.J. Cizek (Ed), *Setting performance standards: Concepts, Methods, and Perspectives*. Erlbaum Mahwah, NJ, pp.427-444.

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. (2001): Conjectures on the rise and call of standard setting: An introduction to context and practice. In G.J. Cizek (Ed), *Setting performance standards: Concepts, Methods, and Perspectives*. Erlbaum Mahwah, N.Jpp. 3-17.

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 Educational 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. 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. (2000a): *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, 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*. Washington, DC, Council of Chief State School Officers, pp. 97-114.

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*. Erlbaum Mahwah, N.J: pp.89-116.

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*. San Francisco, CA, Jossey-Bass, pp. 109-127.

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). Erlbaum Mahwah, N.J, pp. 283-312.

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 Consejería 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.

Jornet J.M. y Backhoff E. (2006) Determinación de NL de los Excale. (Informes del INEE, Septiembre). México, D.F.: INEE.

Jornet J. y Backhoff E. (2006): Niveles de Logro educativos de sexto de primaria y tercero de secundaria en México: Español y Matemáticas: Documento mimeografiado de la Dirección

de Pruebas y Medición. México, D.F.: Instituto Nacional para la Evaluación de la Educación.

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*. Erlbaum Mahwah, N.J, 2001: pp. 53-88.

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*. Erlbaum Mahwah, N.J, 2001: pp. 219-248.

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 behavioural anchoring.* Symposium conducted at the Council of Chief State School Officers National Conference on Large-Scale Assessment, Phoenix, AZ. (June-1996).

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*. Erlbaum Mahwah, N.J: pp.175-218.

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*. Chicago, IL: University of Chicago Press, pp. 83-121.

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*. Erlbaum Mahwah, N.J: pp. 477-485

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 evaluation. *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.* Erlbaum Mahwah, N.J, pp. 249-281.

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 Psychological 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.* Erlbaum Mahwah, N.J: pp. 411-426.

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*. Erlbaum Mahwah, NJ: pp.283-312.

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 multistage 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*. Erlbaum Mahwah, N.J: 2001, pp. 119-157.

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*. Mahwah, NJ: Erlbaum, pp. 159-174.

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.

Ruiz-Primo M.A., Jornet J.M. y Backhoff E. (2006). Acerca de la Validez de los Excale. Colección Cuadernos de Investigación, No. 20. México D.F.: INEE.

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*. Erlbaum Mahwah, N.J, pp. 339-354.

Talente, G., Haist, S. y Wilson, J. (2003): A model for setting performance standards for standardized patient examinations. *Evaluation 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*. Erlbaum Mahwah, N.J, pp. 387-410.

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. pp. 1-38

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*. Erlbaum Mahwah, N.J, pp.19-52.

Zieky, M.J. y Livingston, S. A. (1977): *Basic Skills Assessment. Manual for Setting Standards on the Basic Skills Assessment Tests*. Educational Testing Service, Basic Skills Assessment, Rosedale Road, Princeton, New Jersey.

El equipo técnico responsable del proceso de determinación de NL de los Excale fue conformado por:

Coordinación del proceso:
Dr. Jesús M. Jornet Meliá
Universidad de Valencia
Dr. Eduardo Backhoff Escudero
Director de Pruebas y Medición

Observador / evaluación - validación:
Mtra. Lucía Monroy Cazorla
Investigadora
Mtra. Mª de Lourdes Tanamachi Tanaka
Investigadora

Coordinación de los comités:
Dra. Margarita Peon Zapata
Subdirectora de Español y Ciencias Sociales
Mtro. Andrés Sánchez Moguel
Subdirector de Matemáticas y Ciencias Naturales
Lic. Laura Tayde Prieto López
Coordinadora de Español Primaria
Mtro. Juan Carlos Xique Anaya
Coordinador de Matemáticas Secundaria
Mtra. Cristina Aguilar Ibarra
Coordinadora de Ciencias Naturales
Mtra. Patricia Montero Roa
Coordinadora de Ciencias Sociales
Mtro. Miguel Ángel León Hernández
Coordinador de Matemáticas Primaria
Lic. Ana Laura Villa Blanco*
Coordinadora de Español Secundaria
Mtra. Sara Rivera López
Coordinadora de Español Secundaria

Análisis estadístico:
Fís. Edgar I. Andrade Muñoz
Mtro. en C. José Gustavo Rodríguez Jiménez
Ing. Shaddai Granados Amolitos*

Seminario *Elección del método de determinación de estándares e identificación de puntos de corte* (Método de identificación de NL de los Excale)

Conductor del Seminario:
Dr. Jesús M. Jornet Meliá

Participantes del Consejo Técnico del INEE:
Dr. José Manuel Álvarez Manilla
Presidente del Instituto de Evaluación de Gran Escala (México)
Dr. Eduardo de la Garza Vizcaya
Universidad Autónoma Metropolitana (México)
Dr. Arturo de la Orden Hoz
Universidad Complutense de Madrid (España)
Dr. Guillermo Solano Flores
University of Boulder, Colorado (EUA)

Personal del INEE:
Mtro. Rafael Vidal Uribe*
Mtro. Ricardo Ramírez Aldana
Subdirección de Procesos Estadísticos
Lic. Susana Reyes López
Apoyo técnico de Ciencias Naturales

Invitados, Secretaría de Educación Pública (México):
Mtro. Hugo Balbuena Corro
Dirección General de Desarrollo Curricular
Mtra. Laura Herlinda Lima Muñiz*
Dirección General de Desarrollo Curricular
Dr. Francisco Miranda López*
Coordinador de Asesores de la Subsecretaría de Educación Básica y Normal
Fís. Reyna Estela Silva
Dirección General de Desarrollo Curricular

Otros invitados:
Mtro. Claudio A. Valdivieso Martínez
Universidad Tecnológica de México
Lic. Marcela Arce Tena
Asesora independiente

* Personal que laboró en el INEE o en la SEP durante el proceso de construcción del Modelo de Niveles de Logro Educativo.