Establecimiento de estándares y puntos de corte en exámenes de alto impacto en profesiones de la salud

Laura Delgado Maldonado[a], Melchor Sánchez Mendiola[b,‡,*]

Facultad de Medicina

Resumen

Introducción: Los exámenes sumativos estandarizados de alto impacto en las profesiones de la salud se utilizan para sustentar decisiones de certificación profesional. La definición de estándares y puntos de corte debe ser válida, transparente y defendible, equilibrando la seguridad del paciente, la equidad y la factibilidad.

Objetivo: Sintetizar marcos conceptuales, métodos y consideraciones prácticas para establecer estándares y definir puntos de corte en exámenes de alto impacto, incorporando psicometría, defendibilidad legal y análisis de consecuencias, con énfasis regional latinoamericano.

Método: Revisión narrativa de literatura internacional y regional, documentos técnicos y guías de organismos certificadores. Se comparan enfoques para pruebas de conocimiento (por ejemplo, exámenes de opción múltiple) y de competencias o destrezas (p. ej., simuladores), destacando requisitos, ventajas y limitaciones.

Resultados: Los métodos basados en jueces (Angoff modificado, Ebel) son útiles cuando el contenido está claramente alineado al currículo; Bookmark se apoya en la teoría de respuesta al ítem (TRI) para ordenar ítems y fijar umbrales; Hofstee acota rangos aceptables de aprobación; Beuk y esquemas híbridos ayudan a armonizar el estándar con la dificultad empírica. En evaluación del desempeño, predominan Grupo Limítrofe y Regresión del Limítrofe. Factores críticos: selección y capacitación de panelistas; uso de evidencia empírica (dificultad, discriminación, funcionamiento diferencial de ítems); equiparación entre versiones; estimación del error estándar de medición y bandas de confianza para decisiones de aprobación; documentación y transparencia; monitoreo de impacto y equidad por subgrupos. La defendibilidad legal mejora cuando el estándar se vincula a descriptores de competencia, se sustenta en un argumento de validez y se reporta precisión y consistencia de clasificación.

[a] Consultora independiente, exdirectora General de Medición y Tratamiento de Datos, Instituto Nacional para la Evaluación de la Educación (INEE), Cd. Mx., México.
[b] División de Estudios de Posgrado, Facultad de Medicina, Universidad Nacional Autónoma de México (UNAM), Cd. Mx., México.
ORCID ID:
https://orcid.org/0000-0002-9664-3208
Recibido: 12-septiembre-2025. Aceptado: 18-noviembre-2025.
* Autor para correspondencia: Melchor Sánchez Mendiola.
Correo electrónico: melchorsm@gmail.com, melchorsm@unam.mx

Conclusiones: No existe un método único superior. La elección debe alinearse con el propósito de la evaluación, formato, datos disponibles y recursos. Se recomiendan procesos híbridos, gobernanza clara, trazabilidad documental y evaluación continua de consecuencias para fortalecer la equidad y la confianza pública en los contextos de América Latina.

Palabras clave: *Establecimiento de estándares; puntos de corte; psicometría; certificación médica; exámenes sumativos de alto impacto; América Latina.*

Standard setting and passing scores for high-stakes exams in the health professions

Abstract

Introduction: Highstakes summative examinations in the health professions underpin certification and licensure decisions. Cut scores must be valid, transparent, and defensible, balancing patient safety, fairness, and feasibility. These are needs that are especially relevant across Latin America.

Objective: To synthesize conceptual frameworks, methods, and practical considerations for standard setting and cutscore determination in certification exams, including psychometrics, legal defensibility, and consequences analysis, with a regional emphasis.

Method: Narrative review of international and regional literature, technical reports, and guidelines from certifying bodies. Approaches for knowledge tests (e.g., multiple-choice) and performance assessments (e.g., simulations) are compared, highlighting requirements, strengths, and limitations.

Results: Judge-based methods (modified Angoff, Ebel) are well suited when content is tightly curriculum-aligned; Bookmark leverages item response theory (IRT) to order items and set thresholds; Hofstee constrains acceptable pass/fail ranges; Beuk and hybrid approaches reconcile the standard with empirical difficulty. For performance assessments, Borderline Group and Borderline Regression are predominant. Critical factors include panel selection and training; use of empirical evidence (difficulty, discrimination, differential item functioning); equating across forms; estimating the standard error of measurement and confidence bands for pass/fail decisions; documentation and transparency; and monitoring subgroup impact for equity. Legal defensibility improves when standards are linked to competency descriptors, supported by an explicit validity argument, and accompanied by classification accuracy and consistency evidence.

Implementation in Latin America can benefit from faculty development and capacity building, method selection aligned to data availability (e.g., Angoff/Ebel when IRT calibration is not feasible; Bookmark when item banks exist), robust governance, and routine consequences analyses to ensure fairness and public trust.

Conclusions: No single method is universally superior. Hybrid processes, explicit validity arguments, clear governance, and ongoing psychometric and consequences monitoring strengthen defensibility and equity.

Keywords: *Standard setting; Passing score; Psychometrics; Medical certification; High-stakes summative exams; Latin America.*

"Nunca me propuse ser el mejor.
Es un estándar demasiado bajo.
Me propuse ser bueno".
Henry Mintzberg

INTRODUCCIÓN

Los exámenes sumativos de alto impacto, como son los procesos de certificación de especialidades médicas y otras profesiones de la salud, tienen consecuencias significativas para los candidatos y para la sociedad. La decisión de aprobar o no puede determinar el acceso a ejercer una especialidad, prescribir fármacos, contratación o continuidad laboral. Por ello, los procesos de evaluación deben ser válidos, confiables y legalmente defendibles[1,2]. Un elemento crítico de ese proceso es el establecimiento de estándares (*standard setting*), que conlleva la definición de puntos de corte o calificaciones mínimas aproba-

torias[2-4]. Esta actividad busca determinar el nivel de desempeño que separa a los candidatos competentes de los no competentes, convirtiendo una puntuación continua en una decisión dicotómica (aprobado / no aprobado) o en varios niveles de logro.

El establecimiento de puntos de corte no debe ser arbitrario; debe basarse en un procedimiento racional, colegiado y documentado, y debe reflejar el propósito del examen y las competencias requeridas para la práctica segura[5]. Diversas asociaciones y organismos reguladores han elaborado guías y estándares para asegurar que los procedimientos de establecimiento de estándares sean defendibles y transparentes. Se recomienda vincular los puntos de corte con un análisis de tareas, generar descriptores de niveles de desempeño, seleccionar y capacitar jueces expertos, utilizar métodos psicométricos apropiados y documentar todas las decisiones[4].

Este artículo de revisión presenta un panorama global de los métodos para establecer puntos de corte en exámenes de certificación médica de alto impacto, con énfasis en la situación de América Latina. Se abordan fundamentos conceptuales, métodos basados en la prueba y en los sustentantes, consideraciones psicométricas, análisis de consecuencias, defendibilidad legal y ejemplos de su aplicación en distintos contextos. El objetivo es ofrecer a la comunidad de educación en profesiones de la salud una síntesis sobre este complejo tema.

FUNDAMENTOS CONCEPTUALES

Los exámenes de certificación y recertificación de especialidades médicas se consideran de alto impacto porque sus resultados tienen consecuencias importantes para los candidatos o la sociedad; por ejemplo, la obtención de una licencia profesional o la admisión a un programa de especialidad[6]. Para que un examen de alto impacto sea defendible, deben cumplirse requisitos de validez (medir lo que se pretende medir), confiabilidad (consistencia de los resultados) y equidad. Los *Standards for Educational and Psychological Testing* recomiendan documentar la evidencia de validez y error estándar alrededor de cada punto de corte[7].

Los métodos de establecimiento de estándares se pueden clasificar según la forma en que se interpretan las puntuaciones. Los métodos relativos o normativos comparan a los candidatos entre sí y establecen el punto de corte con base en la distribución de resultados (por ejemplo, aprobar al 70 % con mejores puntuaciones)[2]. En cambio, los métodos absolutos o criteriales usan un criterio externo para determinar la competencia y no dependen del desempeño de los demás aspirantes. Los exámenes de certificación médica requieren idealmente métodos criteriales, ya que la finalidad es asegurar un nivel mínimo de competencia y proteger a los pacientes, independientemente de cuántos candidatos aprueben[4].

MÉTODOS DE ESTABLECIMIENTO DE ESTÁNDARES BASADOS EN LA PRUEBA

Los métodos basados en la prueba (*test-centered*) requieren que un panel de expertos evalúe cada reactivo o tarea en función de su dificultad y pertinencia[8]. Estos métodos no se basan en los resultados de un grupo de candidatos específico, sino en juicios sobre lo que una persona mínimamente competente debería responder correctamente. A continuación, se describen los métodos más comunes.

Método de Angoff y variaciones

El método Angoff es un procedimiento frecuentemente usado en exámenes de opción múltiple. Surgió en 1971 y sigue siendo frecuentemente utilizado debido a su simplicidad y confiabilidad[2,3]. Un panel de 5 a 10 jueces expertos revisa cada reactivo de la prueba e imagina a un "candidato mínimamente competente" (CMC). Para cada pregunta, los jueces estiman la probabilidad de que el CMC la responda correctamente (por ejemplo, 0.70). Las probabilidades se suman para obtener la puntuación de corte en términos de aciertos. Este proceso se realiza de manera individual y luego se discute en grupo; los jueces pueden modificar sus estimaciones hasta alcanzar consenso. Finalmente se promedia la suma de probabilidades para establecer el punto de corte y se puede ajustar mediante equiparación cuando se construyen nuevas versiones del examen.

El método tiene algunas limitaciones. La primera es de carácter conceptual y psicológico. La noción del "candidato mínimo competente" es inherentemente abstracta y ambigua. Cada juez puede construir mentalmente un perfil distinto de este can-

didato, influido por su experiencia profesional, su nivel de exigencia personal y su interpretación de lo que significa "competencia mínima". Esta falta de uniformidad introduce variabilidad que puede impactar negativamente en la consistencia de las estimaciones. En segundo lugar, el método depende en gran medida del juicio experto, lo cual lo hace vulnerable al sesgo cognitivo, la fatiga, y la falta de entrenamiento adecuado. Jueces poco entrenados tienden a sobreestimar o subestimar las probabilidades de éxito del CMC, especialmente cuando los ítems son difíciles o están redactados con ambigüedad. Otra limitación es su escasa base empírica directa, ya que las estimaciones de probabilidad se hacen sin observar el desempeño real de los candidatos. Aunque el Angoff puede complementarse con datos reales a posteriori, por ejemplo, comparando la puntuación de corte con el rendimiento de cohortes anteriores, su núcleo sigue siendo subjetivo. Además, en términos prácticos y logísticos, el método puede volverse oneroso si el número de ítems es muy grande o si se cuenta con un panel de jueces numeroso. La tarea de estimar la probabilidad de éxito en cada pregunta exige tiempo, concentración y habilidades psicométricas mínimas. En contextos con exámenes extensos, esto puede representar un desafío operativo importante.

El método Angoff modificado introduce datos empíricos después de una primera ronda de juicios. Los jueces revisan la dificultad real de cada reactivo en un piloto o examen previo y ajustan sus estimaciones. Esta variante puede mejorar la precisión y la equidad de la puntuación.

Bookmark (marcador)

El método Bookmark es una opción muy utilizada a nivel internacional[3,9,10]. Este procedimiento requiere ordenar los reactivos según su dificultad, usando estadísticas de teoría de respuesta al ítem (TRI), aunque también se han documentado experiencias con uso de métodos clásicos. Los expertos reciben un cuadernillo con los ítems jerarquizados y colocan un "marcador" en el punto donde consideran que un CMC dejaría de responder correctamente. La posición del marcador se traduce en un punto de corte en términos de puntuación; se promedian los marcadores de todos los jueces para obtener la puntuación mínima aprobatoria. Las decisiones dependen de parámetros como la probabilidad elegida (50 o 67 %) de respuesta correcta del CMC. Este método se adapta bien a exámenes adaptativos por computadora y es eficaz cuando se dispone de calibraciones TRI y de un banco de ítems extenso.

Método de Ebel

El método de Ebel es un procedimiento que combina el juicio cualitativo de expertos con un marco cuantitativo basado en la importancia y la dificultad de los ítems. A diferencia de otros métodos puramente empíricos o normativos, Ebel permite analizar cada pregunta del examen en función de su relevancia educativa y su nivel de complejidad, buscando un equilibrio entre exigencia académica y justicia evaluativa. Los jueces organizan los ítems en una matriz de doble entrada con categorías de *importancia* (alta, media, baja) y *dificultad* (fácil, moderada, difícil). Para cada celda, se estima el porcentaje de candidatos mínimamente competentes que deberían responder correctamente los ítems de esa categoría. Posteriormente, cada pregunta se ubica en la celda correspondiente y se le asigna la probabilidad promedio de acierto definida por los jueces. La suma ponderada de todas las probabilidades genera el punto de corte global del examen.

Una de las ventajas del método Ebel es que integra explícitamente la relevancia del contenido con la dificultad técnica, promoviendo deliberaciones más estructuradas y transparentes entre los jueces. Este enfoque facilita la coherencia entre los objetivos educativos y los estándares de desempeño requeridos, y aporta claridad en contextos donde los contenidos poseen diferente peso clínico o educativo, como ocurre en las certificaciones médicas. Sin embargo, el método presenta limitaciones prácticas. Su aplicación requiere tiempo y consenso, especialmente en la clasificación inicial de los ítems, y puede resultar poco operativo en exámenes extensos. Además, la categorización de "importancia" y "dificultad" depende del juicio subjetivo de los panelistas, lo que demanda una capacitación cuidadosa y la documentación del proceso para asegurar la reproducibilidad y la defendibilidad del estándar.

MÉTODOS BASADOS EN LOS SUSTENTANTES

Estos métodos utilizan datos de los candidatos y un criterio externo para determinar el punto de corte. Estos métodos son menos frecuentes en exámenes de opción múltiple, pero se emplean en pruebas de desempeño, como exámenes clínicos objetivos estructurados (ECOE)[3,8].

Grupo limítrofe *(borderline)* y regresión limítrofe

El método del grupo de referencia limítrofe (*borderline group*) es común en ECOE[11,12]. Tras la evaluación, cada examinador asigna una calificación global (por ejemplo, "inferior", "limítrofe" o "superior"). Las listas de cotejo de los candidatos calificados como "limítrofe" se promedian para cada estación; este promedio se convierte en el punto de corte de la estación, y el promedio de todas las estaciones determina el punto de corte global. El método se apoya en la experiencia del evaluador para identificar un desempeño apenas aceptable y produce una puntuación que refleja la dificultad real de cada estación. En una comparación de un examen de certificación en México, el cambio de un punto de corte por criterio (≥6 en una escala de 0 a 10) a un punto de corte determinado por desempeño limítrofe incrementó la tasa de reprobación de los candidatos, mostrando cómo la elección del método puede afectar las decisiones de certificación[13].

El método de regresión limítrofe requiere que los evaluadores asignen tanto una puntuación global como una calificación analítica de cada estación[14]. Se realiza una regresión lineal entre ambas y se determina la puntuación que corresponde al valor "limítrofe" en la escala global (por ejemplo, 2.5 en una escala de 1 a 4). Este valor se utiliza como punto de corte. La ventaja es que integra la calificación y el establecimiento de estándares en un solo proceso y elimina la necesidad de reuniones adicionales; además, ha mostrado estabilidad entre diferentes estaciones y cohortes.

Métodos de grupos contrastantes

El método de grupos contrastantes requiere dos grupos de candidatos con nivel de competencia conocido (por ejemplo, residentes de último año aprobados y estudiantes de pregrado). Se administra el examen a ambos grupos y se identifica el punto de corte que maximiza la clasificación correcta (punto de intersección de las curvas de distribución). Aunque intuitivo, este método depende de contar con grupos representativos y con un criterio externo fiable; en la práctica es poco utilizado en certificaciones de medicina, aunque puede servir para validar otros métodos[15].

MÉTODOS DE COMPROMISO: HOFSTEE Y BEUK

Los métodos de Hofstee y Beuk combinan enfoques criteriales y normativos. En el método de Hofstee, los jueces establecen cuatro valores: la calificación mínima y máxima aceptable como punto de corte y el porcentaje mínimo y máximo de candidatos que deberían aprobar. Estos valores se aplican a la distribución de puntuaciones real para determinar la intersección que define el punto de corte[16,17]. El método se utiliza como moderador para asegurar que la tasa de aprobación no sea irrealmente alta o baja.

El método de Beuk busca equilibrar los juicios de expertos con la tasa de aprobación esperada; los jueces indican el porcentaje mínimo de aciertos y la tasa de aprobación aceptable. Las respuestas de expertos se combinan con la distribución real de resultados para determinar el punto de corte, minimizando discrepancias entre criterios absolutos y realidades normativas[3,18]. El método parte de una estimación de la tasa de aprobación deseada y combina esa expectativa con la distribución real de puntuaciones para ajustar el punto de corte. Su objetivo es equilibrar la exigencia de la prueba con la realidad normativa de los candidatos. Aunque más complejo que Hofstee, puede ser útil en situaciones en las que se desea mantener tasas de aprobación razonables sin comprometer la competencia mínima requerida.

OTROS MÉTODOS Y ENFOQUES HISTÓRICOS

Existen métodos menos utilizados, como el de Cohen (referido a percentiles normativos)[19], y el método arbitrario de 60 % (fijar el punto de corte en 60 % de respuestas correctas)[20]. Estas técnicas suelen carecer de fundamento criterial y pueden generar decisiones inconsistentes. Por esta razón, las organizaciones de

Tabla 1. Comparación de métodos de establecimiento de estándares en exámenes sumativos de alto impacto.

Método	Tipo / Enfoque	Características principales	Ventajas	Desventajas	Sugerencias de uso
Angoff (y Angoff modificado)	Basado en la prueba	Los jueces estiman la probabilidad de acierto del candidato mínimamente competente (CMC) por ítem.	Amplio respaldo empírico; adecuado para opción múltiple.	Demandante para los jueces; subjetivo si hay poca capacitación.	Exámenes de opción múltiple; bancos de ítems estables; paneles de expertos capacitados.
Ebel	Basado en la prueba	Los ítems se clasifican en una matriz según importancia y dificultad.	Estructura sistemática; combina juicio cualitativo y cuantitativo.	Requiere consenso sobre la importancia de los ítems; más complejo de implementar.	Exámenes que cubren múltiples dominios con diferente relevancia.
Bookmark	Basado en la prueba	Ítems ordenados por dificultad (teoría de respuesta al ítem, TRI); el juez coloca un "marcador" en el punto donde el CMC dejaría de responder correctamente.	Integración con bancos de ítems y TRI; útil para escalas y comparaciones longitudinales.	Requiere calibración TRI; menor aplicabilidad en exámenes pequeños.	Certificaciones con bancos grandes; evaluaciones adaptativas por computadora.
Hofstee	Mixto	Los jueces definen límites de calificaciones y tasas mínimas/máximas de aprobación.	Integra juicio experto con realidad empírica; evita extremos.	Más dependiente de la distribución real de puntajes; menos criterial.	Como moderador o verificador tras métodos criteriales (Angoff o Bookmark).
Beuk	Mixto	Combina expectativas de jueces con tasas empíricas de aprobación.	Ajusta criterios absolutos a realidades normativas; útil para estabilidad de tasas.	Requiere datos previos; más complejo de calcular.	Como complemento o ajuste posterior en certificaciones recurrentes.
Grupo limítrofe / Borderline group	Basado en sustentantes	Punto de corte derivado del desempeño promedio de quienes muestran nivel limítrofe.	Representa desempeño real; adecuado para Examen Clínico Objetivo Estructurado (ECOE).	Requiere grandes muestras y consistencia interevaluador.	ECOE y pruebas clínicas.
Regresión limítrofe / Borderline regression	Basado en sustentantes	Se asocia la calificación analítica y la global para estimar el punto limítrofe.	Integración directa en el proceso de evaluación; reproducible.	Requiere análisis estadístico; no siempre aplicable a todas las estaciones.	ECOE con rúbricas analíticas; cohortes medianas o grandes.
Grupos contrastantes	Basado en sustentantes	Punto de intersección entre distribuciones de grupos con competencia conocida.	Intuitivo y empírico.	Difícil obtener grupos representativos; depende del criterio externo.	Validación de otros métodos o estudios de impacto.

certificación modernas privilegian procedimientos basados en la prueba o combinados que tienen respaldo empírico y normativo.

En la **tabla 1** se describen las características principales de los métodos de establecimiento de estándares.

En la **figura 1** se esquematiza el proceso de establecimiento de estándares. Es importante señalar que, en algunos casos como el método de Angoff original, el establecimiento de estándar se realiza antes de tener los resultados del análisis psicométrico de la prueba, por lo que la secuencia de los pasos 5 y 6 de la figura puede variar dependiendo del método específico. El mensaje fundamental es que se deben tomar en cuenta los resultados empíricos en población de sustentantes, para que los jueces puedan calibrarse de forma más aterrizada.

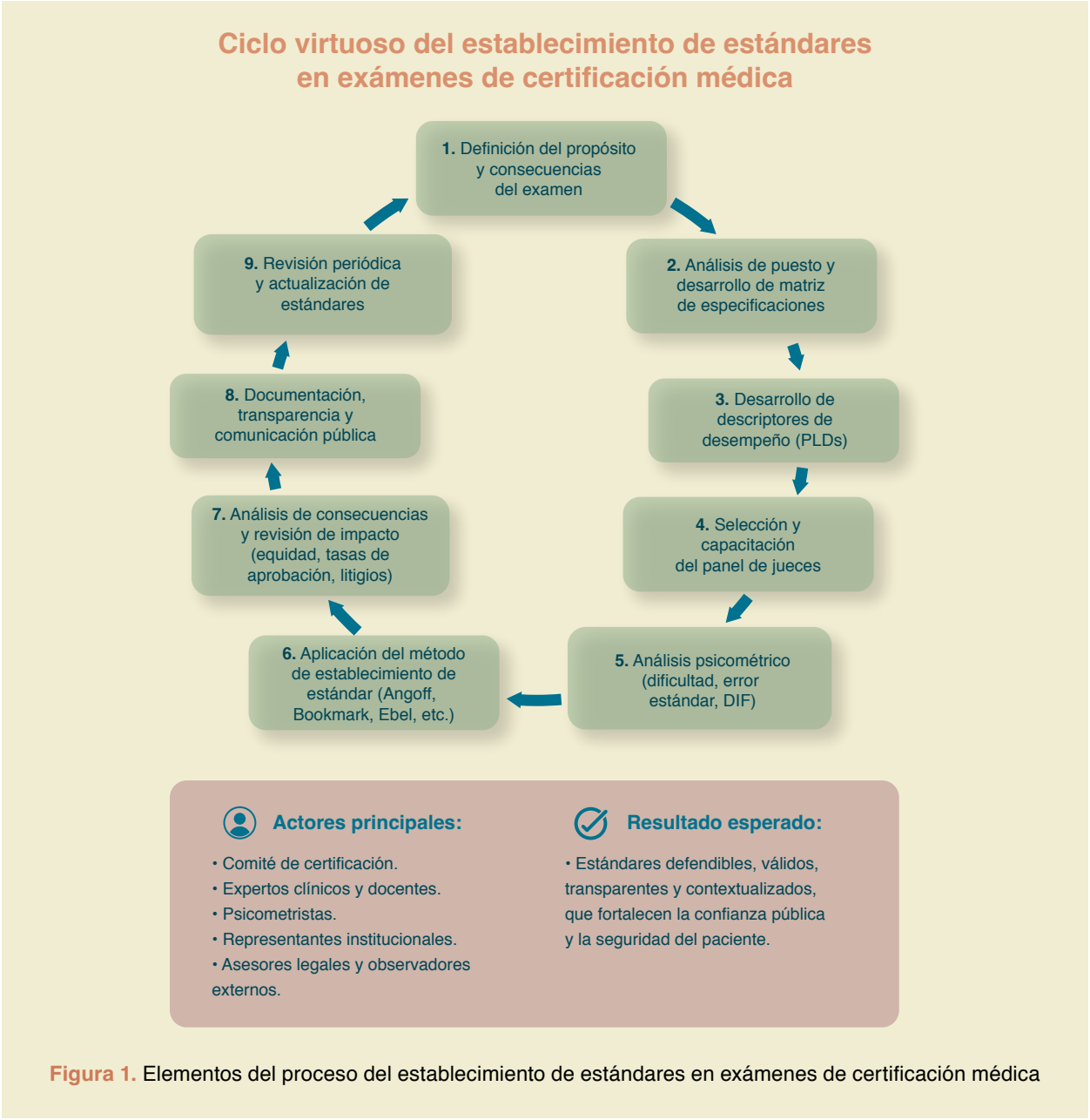

Figura 1. Elementos del proceso del establecimiento de estándares en exámenes de certificación médica

PRINCIPIOS PARA PUNTOS DE CORTE VÁLIDOS Y DEFENDIBLES

Los puntos de corte deben fundamentarse en tres principios: (a) una lógica y procedimiento claros; (b) decisiones colegiadas con un número diverso de expertos; y (c) documentación del proceso para permitir la revisión[2,3,21]. Asimismo, es necesario definir cuántos niveles de desempeño se utilizarán (por ejemplo, no competente, competente y sobresaliente), asignar descriptores claros y conceptualizar lo que significa estar en cada nivel. En el contexto de la educación médica, esto suele implicar describir las tareas clínicas que un *candidato mínimamente competente* debe poder realizar de forma segura y efectiva[8].

La fijación de estándares es un proceso de juicio informado, no una medida matemática estricta: es

Tabla 2. Elementos clave del proceso de establecimiento del punto de corte en exámenes sumativos de alto impacto

Etapa / Elemento	Aspectos críticos	Errores comunes	Recomendaciones prácticas
1. Definición del propósito del examen	Clarificar finalidad (certificación, recertificación, selección).	Métodos no alineados al propósito.	Documentar decisiones y consecuencias esperadas.
2. Análisis de tareas y matriz de especificaciones	Derivar competencias y pesos de contenido.	Subrepresentación de áreas críticas.	Validar con expertos y actualizar cada 3-5 años.
3. Desarrollo de descriptores de niveles de desempeño (PLD)	Describir conductas observables del CMC.	PLD genéricos o ambiguos.	Revisar ejemplos de desempeño real y consensuar verbos conductuales.
4. Selección y capacitación de jueces	Diversidad, experiencia y entrenamiento en el método.	Panel homogéneo o sin capacitación.	5-15 jueces con representación geográfica y de género; sesiones prácticas previas.
5. Análisis psicométrico	Revisión de dificultad, discriminación y error estándar de medición.	No considerar error condicional o DIF.	Incluir reportes de error estándar y análisis de ítems.
6. Aplicación del método	Coherencia interna y registro de deliberaciones.	Falta de documentación o de revisión de datos empíricos.	Registrar estimaciones individuales y discusiones grupales; usar moderadores.
7. Documentación y comunicación	Transparencia del proceso y política de apelación.	Ausencia de evidencia pública o trazabilidad.	Elaborar actas detalladas; publicar criterios generales.
8. Análisis de consecuencias y mejora continua	Impacto en tasas de aprobación, subgrupos y práctica profesional.	No evaluar consecuencias no previstas.	Implementar ciclo de retroalimentación y revisión periódica.

arbitraria pero no caprichosa, pues se fundamenta en evidencia empírica, en el análisis del puesto y en el consenso de jueces[2,8]. De esta manera, el punto de corte se convierte en una decisión educativa y política que debe justificar la clasificación de aspirantes para salvaguardar a los pacientes y al público.

No existe un valor universal o "estándar de oro" que defina la calificación aprobatoria. No hay una cifra perfecta esperando ser descubierta. En realidad, la calificación mínima para aprobar es el resultado del juicio de un grupo de expertos en la materia, quienes la determinan mediante un procedimiento sistemático, reproducible y objetivo. La fortaleza de un estándar defendible radica en aplicar un método riguroso y ordenado para recopilar y analizar las opiniones de los jueces, idealmente sustentado en evidencia científica.

Es importante reconocer que distintos métodos de establecimiento de estándares pueden conducir a diferentes calificaciones aprobatorias; incluso grupos distintos de jueces, aplicando exactamente el mismo procedimiento, pueden llegar a resultados diferentes para una misma evaluación. Estas variaciones no son un problema en sí mismas, salvo si se parte de la premisa errónea de que existe una calificación perfecta o un estándar inmutable. En última instancia, el elemento esencial es el proceso, la forma en que se fundamentan, documentan y comunican las decisiones. Toda calificación aprobatoria implica un acto deliberado, con componentes técnicos y políticos, que refleja juicios de valor necesariamente subjetivos, aunque guiados por criterios de equidad y evidencia. En la **tabla 2** se describen los elementos fundamentales del proceso de establecimiento de punto de corte.

PSICOMETRÍA APLICADA A LA DEFINICIÓN DE PUNTOS DE CORTE

Para que un examen sea defendible, la puntuación de corte debe apoyarse en evidencias de validez (interpretaciones y usos de las puntuaciones) y confiabilidad[4,7]. La validez se documenta demostrando que los ítems están alineados con el análisis de tareas, que los jueces son representativos y que los resultados se correlacionan con otros indicadores de competencia. La confiabilidad se evalúa mediante coeficien-

tes como Cronbach α o KR-20; valores superiores a 0.90 se consideran aceptables para exámenes de alto impacto.

TEORÍA DE RESPUESTA AL ÍTEM Y ANÁLISIS DE DISTRACTORES

La teoría de respuesta al ítem (TRI) permite estimar la dificultad y la discriminación de cada reactivo en una escala común, facilitando la aplicación del método Bookmark y la equiparación de versiones del examen. A partir de los parámetros de dificultad (b) y discriminación (a), se pueden generar cuadernillos ordenados y estimar la probabilidad de respuesta correcta para diferentes niveles de habilidad. La TRI también permite calcular el error estándar condicional de medida, que muestra cómo varía la precisión de la puntuación a lo largo del continuo de habilidades y es esencial para informar sobre la incertidumbre cerca del punto de corte[21].

El análisis de distractores examina cómo los candidatos seleccionan las opciones incorrectas y ayuda a identificar reactivos mal escritos o con claves ambiguas. Los reactivos con índice de discriminación negativo o cercano a cero no aportan información útil y deberían revisarse; un índice de discriminación entre 0.2 y 0.4 se considera moderado, mientras que valores mayores de 0.4 indican excelente discriminación.

El análisis de funcionamiento diferencial del ítem (DIF) evalúa si un reactivo favorece o perjudica injustamente a determinados grupos (por ejemplo, por género o contexto cultural)[22]. Un examen que contiene ítems con DIF significativo puede producir decisiones injustas y vulnerar la defendibilidad legal. Los desarrolladores de exámenes deben realizar análisis DIF y reemplazar o ajustar los reactivos problemáticos. Además, la selección de jueces debe ser diversa, representando distintas regiones, géneros y contextos de práctica para evitar sesgos sistemáticos.

ANÁLISIS DE CONSECUENCIAS Y DEFENDIBILIDAD LEGAL

La definición de un punto de corte no solo tiene implicaciones psicométricas, sino también consecuencias para individuos y para la sociedad. Un falso negativo ocurre cuando un candidato competente falla la prueba; puede generar pérdida de oportunidades, litigios y afecta la moral de los aspirantes. Un falso positivo implica aprobar a un candidato no competente, exponiendo a pacientes a riesgos y afectando la reputación de la certificación[1,2,23]. Los comités de certificación deben buscar un equilibrio entre estos riesgos y considerar la finalidad protectora del examen.

Un examen de alto impacto puede ser impugnado judicialmente por candidatos reprobados o por terceros si no se cumplen principios de equidad y debido proceso. La defendibilidad legal se fortalece cuando se siguen estándares reconocidos, se documentan los procedimientos, se emplean métodos de establecimiento de estándares justificados y se realiza un análisis continuo de equidad y validez[4,7,24]. Las buenas prácticas incluyen realizar un análisis de tareas exhaustivo, involucrar a un panel diverso de expertos en todas las fases, utilizar análisis psicométricos para calibrar y revisar ítems, documentar las deliberaciones de los jueces y comunicar claramente la política de reexaminación y apelación. Los estándares internacionales (AERA-APA-NCME) requieren que se informe el error estándar en la vecindad de cada punto de corte y que se provea información sobre la fiabilidad de las decisiones, no solo de las puntuaciones[7].

SESGOS Y EQUIDAD EN AMÉRICA LATINA

En la región latinoamericana, existen desafíos adicionales relacionados con la diversidad de contextos educativos, diferencias en recursos y aspectos socioculturales. Muchos exámenes de certificación se importan o adaptan de modelos extranjeros, lo que puede generar problemas de ajuste cultural y técnico. Por ejemplo, algunas organizaciones utilizan métodos Angoff y Bookmark pero enfatizan la capacitación de jueces y la documentación del proceso para asegurar que los puntos de corte reflejen el contexto mexicano[25]. Los consejos de especialidades médicas han adoptado variaciones locales: el Consejo Mexicano de Medicina Interna aplica el método Angoff y luego transforma las puntuaciones a una escala que incluye categorías de excelencia (https://www.cmmi.org.mx/regulacion/docs/Proceso_de_Examen_CMMI_firmado.pdf), mientras que otros, como el Consejo Mexicano de Reumatología, han experimentado con métodos de desempe-

ño limítrofe[13]. Estos casos muestran la necesidad de adaptar los procedimientos a las características de cada profesión y población, y de evaluar el impacto en la equidad.

RECOMENDACIONES PARA IMPLEMENTAR UN PROCESO DE ESTABLECIMIENTO DE ESTÁNDARES EN EXÁMENES DE CERTIFICACIÓN MÉDICA

1. **Definir el propósito del examen y las consecuencias asociadas.** Es imprescindible establecer si el examen se utilizará para otorgar certificación inicial, recertificación, acreditación de competencias específicas u otros fines. La definición guiará la selección del método de establecimiento de estándares y el número de niveles de desempeño.
2. **Realizar un análisis de tareas y desarrollar una matriz de especificaciones.** La matriz debe reflejar las competencias esenciales y los pesos relativos que cada área tendrá en el examen. Esto asegura que los reactivos representen adecuadamente el dominio y que el punto de corte sea coherente con las tareas reales de la práctica clínica.
3. **Redactar descriptores de desempeño para cada nivel.** Estos descriptores ayudan a los jueces a conceptualizar al CMC y sirven para comunicar a los candidatos lo que se espera de ellos. El desarrollo de descriptores de nivel de desempeño (PLDs) es un paso fundamental.
4. **Seleccionar y capacitar a los jueces.** Se recomienda incluir entre 5 y 15 expertos con experiencia clínica y docente, representativos de la diversidad geográfica, de género y de práctica. Deben recibir capacitación sobre el método que se empleará, practicar con reactivos de ejemplo, discutir la definición del CMC y familiarizarse con los criterios de puntuación. La capacitación incluye explicar los principios de validez, confiabilidad y equidad, así como evitar sesgos cognitivos.
5. **Elegir el método de establecimiento de estándar.** Para exámenes de opción múltiple con grandes bancos de ítems, el método Angoff modificado o el Bookmark son apropiados. Para exámenes de habilidades clínicas (ECOE), los métodos de grupo limítrofe o de regresión limítrofe son preferidos. Los métodos combinados (Hofstee, Beuk) pueden servir como complemento o moderador cuando se desea considerar la tasa de aprobación esperada.
6. **Aplicar el procedimiento y documentar cada paso.** Se debe registrar la discusión de los jueces, las justificaciones para cada estimación, los cambios realizados tras revisar datos empíricos y las decisiones finales. Esta documentación es esencial para la defendibilidad legal.
7. **Analizar las consecuencias y retroalimentar el proceso.** Después de aplicar el examen, se debe analizar la distribución de resultados, la relación entre el punto de corte y la tasa de aprobación, el error estándar alrededor del punto de corte y el impacto sobre distintos grupos demográficos. Este análisis permite ajustar el punto de corte, revisar ítems problemáticos y mejorar la equidad.
8. **Revisar y actualizar los estándares periódicamente.** Los cambios en la práctica médica, avances tecnológicos o nuevas regulaciones pueden requerir actualizar las competencias y, por ende, los puntos de corte. Las guías recomiendan revisar los estándares cada 3 a 5 años o cuando se realiza una nueva actualización del análisis de puesto.

PERSPECTIVAS FUTURAS Y DESAFÍOS

Aunque las metodologías de establecimiento de estándares están bien documentadas, existen áreas que requieren mayor atención. Una línea emergente es el uso de modelos de cómputo adaptativos y aprendizaje automático para apoyar la toma de decisiones[26]. Los algoritmos de inteligencia artificial pueden ayudar a priorizar ítems que mejor discriminan entre niveles de competencia, optimizar la selección de reactivos durante las pruebas adaptativas y simular el impacto de distintos puntos de corte. Sin embargo, su implementación en certificaciones médicas plantea retos éticos y de transparencia, pues la lógica de los algoritmos debe ser comprensible para los jueces y auditada por expertos.

Otro desafío es la adaptación cultural de los exámenes. Muchos bancos de reactivos se derivan de contextos anglosajones; traducirlos literalmente puede introducir sesgos y reducir la validez local. Se requieren estudios de validación transcultural, revisión lingüística y consulta con especialistas lo-

cales para asegurar que las competencias evaluadas reflejan la práctica regional y que los pacientes y escenarios de las preguntas son representativos.

En términos de investigación, son escasos los estudios que evalúen la efectividad de distintos métodos en América Latina. Se necesitan comparaciones empíricas entre el Angoff, Bookmark, Hofstee y los métodos de desempeño limítrofe en certificaciones reales, evaluando no solo las tasas de aprobación sino también la relación con indicadores de desempeño clínico posterior y la satisfacción de los candidatos. Asimismo, conviene explorar la percepción de justicia y transparencia entre aspirantes y empleadores para fortalecer la confianza social en estos procesos. La cooperación entre consejos de especialidades y universidades de la región facilitará la creación de bancos de ítems, la capacitación de jueces y el intercambio de mejores prácticas.

CONCLUSIONES

El establecimiento de estándares y puntos de corte en exámenes sumativos de alto impacto es un proceso complejo que combina juicio experto, técnicas psicométricas y consideraciones éticas y legales. Los métodos basados en la prueba, como el Angoff modificado y el Bookmark, son ampliamente utilizados en exámenes de opción múltiple y cuentan con respaldo teórico y empírico. Los métodos basados en sustentantes, como el grupo limítrofe y la regresión limítrofe, son útiles en evaluaciones de habilidades clínicas. Los métodos combinados, como Hofstee y Beuk, permiten equilibrar exigencias absolutas y tasas de aprobación aceptables, pero requieren manejo cuidadoso.

La literatura enfatiza que el punto de corte debe asociarse con una competencia mínima segura, no con un porcentaje arbitrario de aciertos, y que la decisión final debe considerar las consecuencias sociales de aprobar o reprobar candidatos. La defendibilidad legal se sustenta en seguir procedimientos establecidos, capacitar a los jueces, documentar las deliberaciones y demostrar que el examen es válido, confiable y equitativo.

En América Latina, los organismos de certificación han adoptado estas metodologías con adaptaciones locales. Sin embargo, la región necesita más investigación empírica sobre la aplicación de estas técnicas, la percepción de los candidatos y los resultados en la práctica clínica. Un enfoque continuo de mejora y la transparencia en la comunicación con la sociedad fortalecerán la confianza en los exámenes de certificación y, en última instancia, contribuirán a la seguridad de los pacientes y a la calidad de la atención en salud.

REFERENCIAS

1. Norcini JJ, Lipner RS, Grosso LJ. Assessment in the Context of Licensure and Certification. Teach Learn Med. 2013;25(suppl 1):S62-7. https://doi.org/10.1080/10401334.2013.842909
2. Norcini JJ. Setting standards on educational tests. Med Educ. 2003;37(5):464-9. https://doi.org/10.1046/j.1365-2923.2003.01495.x
3. Cizek, Gregory J., and Michael B. Bunch. Standard Setting. Thousand Oaks, CA: SAGE Publications, Inc.; 2007. https://doi.org/10.4135/9781412985918
4. Davis-Becker, S., & Buckendahl, C.W. (Eds.). Testing in the Professions: Credentialing Policies and Practice. 1st ed. Routledge; 2017. https://doi.org/10.4324/9781315751672
5. Hejri SM, Jalili M. Standard setting in medical education: fundamental concepts and emerging challenges. Meed J Islam Repub Iran. 2014;28:34. Disponible en: https://tinyurl.com/37c5jzpr
6. Sánchez-Mendiola M, Delgado-Maldonado L. Exámenes de alto impacto: implicaciones educativas. Inv Ed Med. 2017;6(21):52-62. https://doi.org/10.1016/j.riem.2016.12.001
7. American Educational Research Association, American Psychological Association, & National Council on Measurement in Education (Eds.). Standards for educational and psychological testing. American Educational Research Association. 2014. Disponible en: https://www.testingstandards.net/open-access-files.html
8. Yudkowsky R, Park YS, Downing SM. (Eds.). Assessment in Health Professions Education (2nd ed.). Routledge. 2019. https://doi.org/10.4324/9781138054394
9. Afrashteh MY. Comparison of the validity of bookmark and Angoff standard setting methods in medical performance tests. BMC Med Educ. 2021;21(1):1. https://doi.org/10.1186/s12909-020-02436-3
10. Lypson ML, Downing SM, Gruppen LD, Yudkowsky R. Applying the Bookmark method to medical education: standard setting for an aseptic technique station. Med Teach. 2013;35(7):581-5. https://doi.org/10.3109/0142159X.2013.778395
11. Kaufman DM, Mann KV, Muijtjens AMM, Vleuten CPM van der. A comparison of standard-setting procedures for an OSCE in undergraduate medical education. Acad Med. 2000;75(3):267-71. https://doi.org/10.1097/00001888-200003000-00018
12. Moreno-López R, Hope D. Can borderline regression method be used to standard set OSCEs in small cohorts? Eur J Dent Educ. 2022;26(4):686-91. https://doi.org/10.1111/eje.12747

13. Pascual-Ramos V, Bernard-Medina AG, Flores-Alvarado DE, Portela-Hernández M, Maldonado-Velázquez M del R, Jara-Quezada LJ, et al. El método para establecer el punto de corte en el examen clínico objetivo estructurado define el desempeño de los candidatos a la certificación como reumatólogo. Reum Clínica. 2018;14(3):137-41. https://doi.org/10.1016/j.reuma.2016.11.007
14. Hejri SM, Jalili M, Muijtjens AMM, Vleuten CPMVD. Assessing the reliability of the borderline regression method as a standard setting procedure for objective structured clinical examination. J Res Med Sci: Off J Isfahan Univ Med Sci. 2013;18(10):887-91. Disponible en: https://pmc.ncbi.nlm.nih.gov/articles/PMC3897074/pdf/JRMS-18-887.pdf
15. Jørgensen M, Konge L, Subhi Y. Contrasting groups' standard setting for consequences analysis in validity studies: reporting considerations. Adv Simul. 2018;3(1):5. https://doi.org/10.1186/s41077-018-0064-7
16. Burr SA, Whittle J, Fairclough LC, Coombes L, Todd I. Modifying Hofstee standard setting for assessments that vary in difficulty, and to determine boundaries for different levels of achievement. BMC Med Educ. 2016;16(1):34. https://doi.org/10.1186/s12909-016-0555-y
17. Khan U. Standard setting OSCE: A comparison of arbitrary and Hofstee methods in a low stake OSCE. Asia Pac Sch. 2024;9(3):15-21. https://doi.org/10.29060/TAPS.2024-9-3/OA3129
18. Wyse AE. A Critical Look into the Beuk Standard-Setting Method. Educational Measurement: Issues and Practice, 2020; 39: 52-60. https://doi.org/10.1111/emip.12317
19. Taylor CA. Development of a modified Cohen method of standard setting. Med Teach. 2011;33(12):e678-82. https://doi.org/10.3109/0142159X.2011.611192
20. Yousef MK, Alshawwa L, Tekian A, Park YS. Challenging the arbitrary cutoff score of 60%: Standard setting evidence from preclinical Operative Dentistry course. Med Teach. 2017;39(sup1):S75-9. https://doi.org/10.1080/0142159X.2016.1254752
21. Lahner FM, Schauber S, Lörwald AC, Kropf R, Guttormsen S, Fischer MR, et al. Measurement precision at the cut score in medical multiple choice exams: Theory matters. Perspect Med Educ. 2020;9(4):220-8. https://doi.org/10.1007/s40037-020-00586-0
22. Hope D, Adamson K, McManus IC, Chis L, Elder A. Using differential item functioning to evaluate potential bias in a high stakes postgraduate knowledge based assessment. BMC Med Educ. 2018;18(1):64. https://doi.org/10.1186/s12909-018-1143-0
23. Searle J. Defining competency - the role of standard setting. Med Educ. 2000;34(5):363-6. https://doi.org/10.1046/j.1365-2923.2000.00690.x
24. Norcini J, Anderson MB, Bollela V, Burch V, Costa MJ, Duvivier R, Hays R, Palacios Mackay MF, Roberts T, Swanson D. 2018 Consensus framework for good assessment. Med Teach. 2018;40(11):1102-1109. https://doi.org/10.1080/0142159X.2018.1500016
25. Ceneval Comunica. (2022, 4 de abril). Puntos de corte en las pruebas con referente de calificación criterial. Ceneval Comunica, Boletín 28. Disponible en: https://ceneval.edu.mx/blog/2022/04/04/puntos-de-corte-en-las-pruebas-con-referente-de-calificacion-criterial/
26. Hao J, von Davier AA, Yaneva V, Lottridge S, von Davier M, Harris DJ. Transforming Assessment: The Impacts and Implications of Large Language Models and Generative AI. Educational Measurement: Issues and Practice. 2024;43:16-29. https://doi.org/10.1111/emip.12602