CORRECCIÓN DEL EUNACOM MEDIANTE TEORÍA DE RESPUESTA AL ÍTEM (IRT)

Dr. Beltrán Mena[1], Víctor Pedrero, Ph.D[2]

INTRODUCCIÓN

A partir del examen de diciembre de 2024, la corrección del EUNACOM se realizará utilizando la *Teoría de Respuesta al Ítem* (*IRT*, por su sigla en inglés). Este enfoque permite determinar la dificultad de las preguntas (ítems), de manera independiente del conocimiento del grupo de personas que las responden. Esta teoría puede aplicarse a través de varios modelos (EUNACOM utilizará el modelo de Rasch) y constituye el estándar actual para exámenes de alta consecuencia, por lo que es utilizada en los exámenes de habilitación médica en países desarrollados, así como en la PAES y el SIMCE en nuestro país.

Para la aplicación del método, la calibración del banco de preguntas y la verificación de sus resultados EUNACOM trabajó entre 2020 y 2022 con la asesoría del *National Board of Medical Examiners (NBME)*, entidad responsable de elaborar y corregir los exámenes de medicina usados en Estados Unidos, con quienes se firmó un convenio de colaboración técnica para equiparar los métodos de corrección de ambas entidades. El trabajo fue supervisado en todo momento por el Consejo Estadístico del EUNACOM. Durante este período y con el fin de afinar el proceso, la corrección se realizó con IRT paralelamente a la corrección tradicional, encontrándose hoy el equipo técnico en condiciones de aplicar el nuevo método. El nuevo procedimiento fue aprobado por el Consejo de Decanos de ASOFAMECH el lunes 27 de mayo de 2024.

El nuevo método presenta grandes ventajas respecto al anterior: alinea la corrección del EUNACOM con estándares internacionales, garantiza la equidad en concursos y permite el seguimiento en el tiempo de cambios curriculares por parte de las escuelas.

El modelo IRT, sin embargo es más complejo, requiere ser aplicado por profesionales especializados, utiliza software dedicado y es más difícil de explicar a no especialistas. Quienes estén familiarizados con psicometría y los conceptos mencionados pueden ir directo a la hoja técnica de corrección, donde se detalla el algoritmo y fórmulas utilizados.

El presente documento está dirigido a profesores y estudiantes de medicina y pretende explicar en forma intuitiva los fundamentos y ventajas de IRT respecto al sistema actual, así como detallar el proceso con que se corrige y puntúa el EUNACOM.

(1) Director del EUNACOM
(2) Jefe Unidad Técnica del EUNACOM

¿POR QUÉ CAMBIAR DE MÉTODO?

El método simple utilizado hasta ahora en la corrección del EUNACOM fundamentalmente contabiliza las respuestas correctas y las transforma a una escala proporcional de puntuación con valores entre cero y 100. Además, con el fin de monitorear la calidad del examen el EUNACOM utiliza indicadores derivados de la *Teoría Clásica de la Medición (TCM)* tales como error estándar, confiabilidad, correlaciones biseriales y otros. Un método como este es más que suficiente para puntuar en forma justa y para comparar los conocimientos de quienes rindieron un mismo examen, sin embargo resulta poco confiable para comparar puntajes de personas que rindieron exámenes distintos o para comparar los resultados de una escuela a lo largo del tiempo.

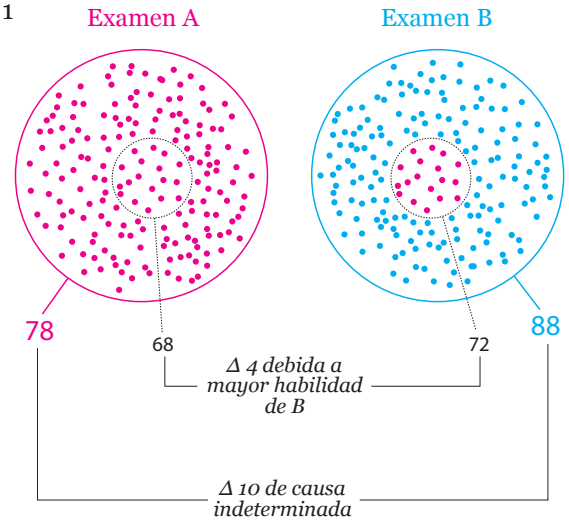
Así, por ejemplo, si dos egresados de años consecutivos que rindieron distintos exámenes y obtuvieron uno 75 puntos y el otro 79, postulan a un mismo concurso de especialidad, ¿cómo saber si el examinado que obtuvo 4 puntos más posee en realidad 4 puntos más de conocimiento o tuvo la suerte de rendir un examen 4 puntos más fácil? La duda será la misma cuando una escuela pretenda evaluar el impacto de mejoras curriculares a lo largo del tiempo, ya que no tiene cómo distinguir si la variación de puntaje entre dos promociones refleja cambios en el aprendizaje o solo diferencias en la dificultad del examen.

Para resolver esta duda sería necesario poder medir la dificultad de un examen (el promedio de dificultad de las preguntas que lo componen), pero en *TCM* la dificultad de una pregunta (p) se define simplemente como la proporción de personas que la respondió correctamente, que evidentemente depende de la habilidad (conocimiento) del grupo que la responda.

Existen procedimientos, conocidos como *equating*, que intentan resolver este problema, pero el "punto ciego" de la *TCM* –no poder separar dificultad de habilidad– hace difícil aplicarlos sin asumir supuestos riesgosos.

El concepto que subyace a los métodos de equating puede ilustrarse con el ejemplo de la Fig 1: La promoción A rinde el examen A, obteniendo un puntaje promedio de 78 y la promoción B rinde el examen B, obteniendo un puntaje promedio de 88. Cada punto de la figura representa una de las 180 preguntas del examen. Para poder saber a qué atribuir la diferencia de 10 puntos a favor de la promoción B, se diseñó

Fig 1



el examen B con *equating* en mente, para lo cual se tuvo cuidado de incluir en el examen B 18 preguntas ya utilizadas en el examen A (preguntas repetidas).

Inicialmente se corrigen sólo las preguntas repetidas, observando que los puntajes son 68 para A y 72 para B. Como las preguntas son las mismas, suponemos que su dificultad es idéntica y atribuimos los 4 puntos de diferencia a mayor conocimiento en la promoción B.

Ahora sabemos que 4 de los 10 puntos extra obtenidos por la promoción B se debían en efecto a mayor conocimiento en la promoción B y el resto (6 pts) debe atribuirse a que las preguntas eran más fáciles.

Restamos entonces 6 pts al puntaje general de la promoción B y obtenemos un puntaje corregido de 82 puntos para esa promoción.

Los supuestos descritos van acumulando un grado desconocido de error en el ajuste de puntajes. Cuando las consecuencias de los resultados no son críticas, este error de magnitud desconocida puede ser aceptable, pero en el caso de EUNACOM el asignar igual puntaje a habilidades distintas tiene altas consecuencias:

a) El puntaje de aprobación (51 pts) puede no reflejar idéntica habilidad en cada examen. Es posible que dos personas con conocimientos similares sean una aprobada y la otra reprobada si rinden exámenes distintos.

b) Iguales puntajes informados pueden no reflejar idéntica habilidad en cada examen. Es posible que una persona con menos conocimientos que otra sea favorecida en un concurso si rindieron examenes diferentes.

c) Intervenciones curriculares importantes pueden no verse reflejadas en los resultados, o peor aún, los resultados pueden ocultar un deterioro en la formación.

Volvamos al ejemplo de la Fig 1. El supuesto teórico de que el grupo de 18 preguntas repetidas conserva su dificultad de un uso a otro es cuestionable. Existen varias situaciones que pueden hacer variar la dificultad de una pregunta entre dos aplicaciones. La más importante es su *exposición* (circulación entre estudiantes), lo que hará que las respuestas a preguntas usadas en el examen A sean conocidas por quienes responden el examen B, que los hará aparecer como poseedores de mayor conocimiento, distorsionando el ajuste de las preguntas no repetidas. Como solución se podrían comparar las dificultades de las 18 preguntas repetidas entre el examen A y el examen B antes de realizar el *equating* y no utilizar aquellas cuya dificultad haya disminuído, atribuyendo esta variación a exposición de la pregunta. Lamentablemente no podremos saber si la variación se debió a que en efecto la pregunta circuló pero los examinados saben lo mismo, o a que no circuló y en efecto los examinados saben más. La debilidad principal de la *TCM* –no poder separar la dificultad de una pregunta de la habilidad de quienes la respondieron– nos impide comparar de manera confiable exámenes distintos rendidos por grupos distintos.

Para un uso justo del examen debemos aplicar algún método de *equating*, pero antes necesitamos determinar la dificultad de una pregunta de manera independiente de la habilidad de quienes la respondan.

Aquí es donde surge la utilidad de la *Teoría de Respuesta al Ítem (IRT)*, perfeccionada por varios matemáticos desde 1943 y aceptándose como estándar en psicometría desde que la publicaran en forma de libro Lord y Novick en 1968[1]. La aparición posterior de distintos programas computacionales permitió ponerla en práctica, utilizándose como estándar en examenes de alta consecuencia.

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

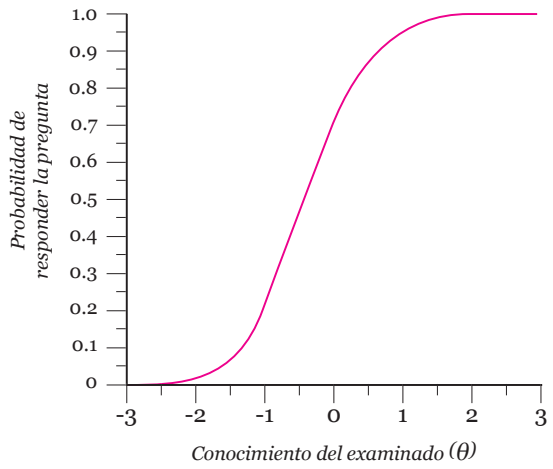
TEORÍA DE RESPUESTA AL ÍTEM (IRT)

IRT determina la dificultad de las preguntas y la habilidad de los examinados en forma independiente, lo que permite realizar *equating* basado únicamente en la dificultad del examen, sin que este se confunda con el conocimiento de quienes lo rindieron.

La habilidad de un examinado se representa con el símbolo θ (theta) y corresponde al aspecto que se pretende medir, inaccesible a la medición directa; en el caso de EUNACOM esta habilidad son los *conocimientos de medicina general* y en este artículo nos referiremos a ella indistintamente como *habilidad* o *conocimientos*.

Para separar conocimiento de dificultad, IRT no calcula un parámetro de dificultad para una pregunta, si no que determina una curva completa que describe la probabilidad de escoger la respuesta correcta dado cualquier nivel de habilidad del examinado que enfrente esa pregunta. Esta curva se conoce como la *curva característica* de una pregunta (Fig 2).

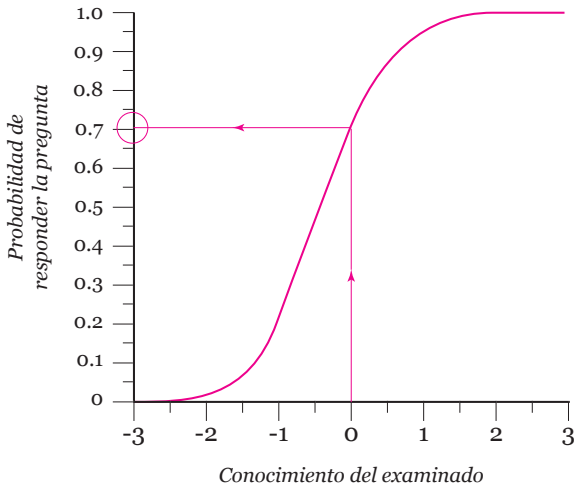
Fig 2



La curva característica de cada pregunta se calcula en base a las respuestas y resultados de una promoción en el examen completo. Cómo no se sabe cuál es el máximo ni el mínimo de conocimiento posible, se define que el centro de la escala sea un valor cero, que corresponde a algún nivel de habilidad absoluta, con unidades positivas hacia la derecha por sobre ese nivel de habilidad y negativas hacia la izquierda, por debajo del nivel cero de habilidad. La curva utilizada para modelar esta probabilidad corresponde a la llamada *función logística*, ampliamente utilizada en modelamiento de fenómenos naturales, que predice con gran precisión los datos obtenidos en respuesta a exámenes.

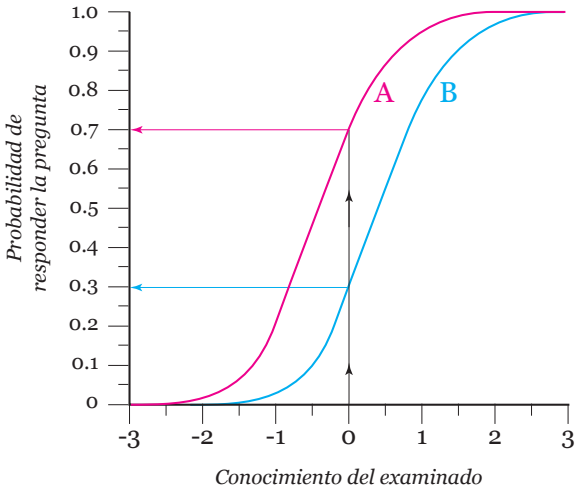
La función representa un fenómeno en que el crecimiento de lo observado (en este caso probabilidad de responder la pregunta) depende de la densidad de elementos requeridos para hacer crecer el valor (en este caso conocimientos clínicos). En el ejemplo de la Fig. 3 un examinado de conocimiento cero *logits* tiene una probabilidad de 0.7 de responder correctamente la pregunta graficada.

Fig 3



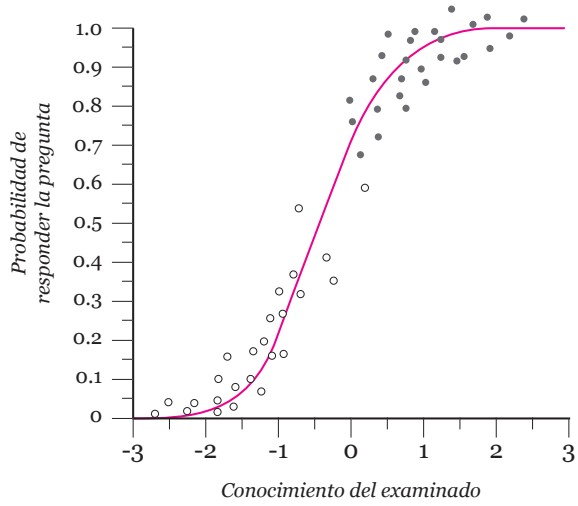
En la Fig 4 podemos ver que el mismo examinado con nivel de conocimientos cero *logits*, tiene un 70% de probabilidades de responder correctamente la pregunta A, pero sólo un 30% de probabilidades de responder correctamente la pregunta B, que es por lo tanto, más difícil.

Fig 4



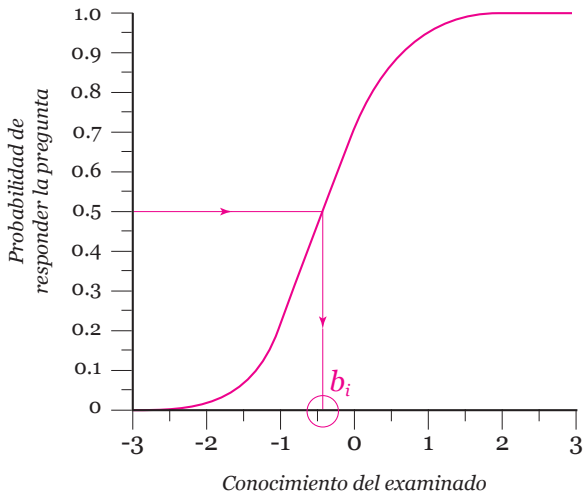
Para posicionar la curva, el proceso ensaya todas sus posibles posiciones hasta encontrar aquella que mejor explique los datos observados. En el modelo utilizado por EUNACOM (modelo de Rasch) la curva se desplaza hacia derecha o izquierda, pero no cambia su forma, esto significa que, una vez posicionada en la escala de *logits*, la curva describirá la dificultad de la pregunta para toda la gama de conocimientos, independientemente de la habilidad de la población que se usó para construirla. Esta propiedad se denomina *invarianza de medición* y puede intuirse en la Fig. 5, donde se ve cómo la curva se puede construir tanto con una población de baja habilidad (puntos blancos), como con una de alta habilidad (puntos grises). Habrá sólo una curva que maximice el ajuste de la curva con los datos observados.

Fig 5



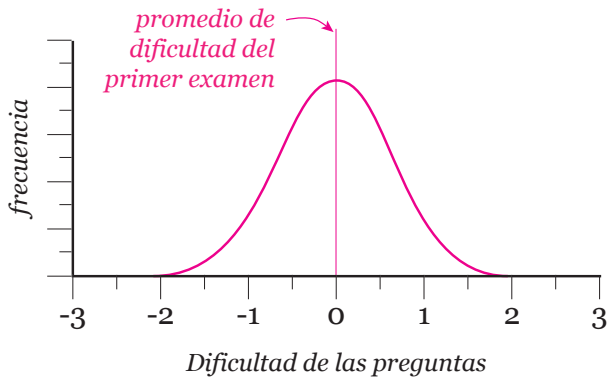
Para contar con un valor de dificultad de la pregunta, que nos permita compararla con otras o determinar la dificultad promedio de un examen completo, se define la dificultad de una pregunta (b_i) como el conocimiento necesario para tener un 50% de probabilidades de responderla correctamente. En el ejemplo de la Fig 6 esta dificultad resulta ser aproximadamente -0.45, lo que significa que personas con habilidad -0.45 logits la responderán correctamente con un 50% de probabilidad.

Fig 6



Construidas las curvas características de cada pregunta, el sistema obtiene la dificultad (b_i) de cada una, lo que permite obtener el promedio de sus dificultades, que corresponde a la dificultad del examen y que el modelo centra en cero (ver Fig 7).

Fig 7



Obtenemos así la dificultad de un primer examen, que será nuestro examen de referencia y que llamaremos Examen R. Las escalas de todos los futuros exámenes deberán ser corregidas para llevarlas a la misma escala de este examen de referencia.

Al corregir el siguiente examen (Examen B), obtendremos resultados expresados en una escala similar a la del Examen R, con la salvedad de que los ceros de cada examen representarán dificultades distintas, ya que siendo distintas las preguntas de ambos exámenes, su dificultad será necesariamente distinta. Para averiguar en cuánto debemos desplazar la escala del Examen B para que su cero coincida en un mismo nivel de dificultad con el cero del Examen R debemos aplicar equating:

1.- Escogemos una muestra de preguntas representativas del Examen R, cuyo promedio de dificultad corresponderá a la del examen completo.

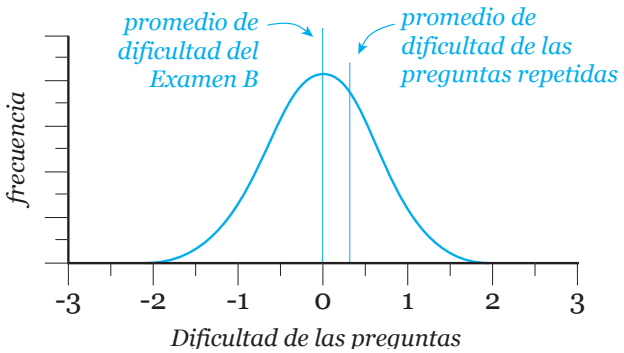
2.- Incluímos esta muestra de preguntas en el nuevo examen (Examen B) y completamos el resto del examen con preguntas nuevas.

3.- Aplicamos el Examen B y calculamos las curvas características de cada una de sus preguntas, incluídas las repetidas. El promedio de dificultad de las preguntas de este nuevo examen será por definición de cero logits, pero sabemos que éste cero no representa el mismo nivel de dificultad que el del examen R. Debemos averiguar ahora la magnitud de esa diferencia.

4.- Para ello separamos del Examen B las preguntas repetidas y comparamos la distribución de sus dificultades, que debe ser igual a la que tenían cuando se usaron en el Examen R. Cualquier variación en su dificultad más allá de un límite tolerable implica que la pregunta se expuso (circuló) y se elimina del grupo de equating. Con IRT podemos hacerlo sin riesgo de confundir dificultades y habilidades. Procedemos ahora al equating basados en un grupo de preguntas repetidas que conservan su dificultad.

5.- La dificultad promedio de las preguntas repetidas en el Examen B no tiene por qué ser de cero logits como el promedio del total del Examen B. En el ejemplo de la Fig 8 las preguntas repetidas resultan tener una dificultad 0.3 logits mayor que la del total del examen, pero sabemos que en el Examen R tenían un valor de cero. Así averiguamos que el cero del Examen B se desplazó 0.3 respecto al cero del Examen R.

Fig 8



6.- Restamos 0.3 logits a cada una de las preguntas del Examen B, con lo cual la dificultad representada por su cero y por toda su escala coincidirá con la del Examen R y podremos comparar todos los resultados de ambos exámenes con una misma escala de referencia.

7.- Al contar con una misma escala en ambos exámenes, podemos saber qué tanto más difícil o más fácil resultó el Examen B respecto al Examen R y usar ese dato para corregir el puntaje obtenido por cada examinado del Examen B, de manera que pueda compararse de manera justa con el Examen R.

8.- Esto significa que el puntaje de cada examinado no dependerá sólo de sus respuestas correctas, si no también de la dificultad del examen que le tocó responder. Así, si un examinado rindió un examen más difícil que el examen de referencia,

requerirá menos respuestas correctas para demostrar el mismo nivel de habilidad que quien rindió el Examen R. La ecuación específica puede verse en la *hoja técnica de corrección*.

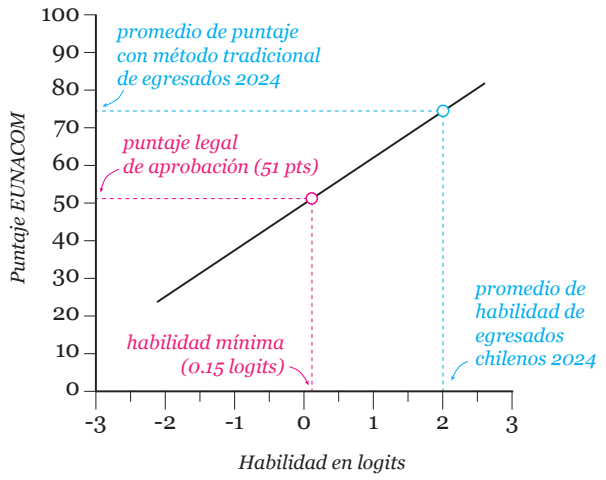
9.- Una vez determinada la habilidad de cada examinado (expresada en logits), este valor se convierte en puntaje EUNACOM mediante una transformación lineal. Cualquier recta puede cumplir esta función, a condición de que se conserve igual en futuros exámenes. Sin embargo al trazarla se prefirió definir dos puntajes relevantes:

a) El puntaje de aprobación (51 pt) corresponderá a 0.1 logits de conocimientos [1].

b) El promedio de conocimientos (en logits) de los examinados de diciembre 2024 corresponderá al promedio de puntaje EUNACOM que hubiesen obtenido con la corrección tradicional. Esto asegura que al momento de su puesta en marcha no haya diferencia promedio entre ambos métodos.

La ecuación de conversión resultante sólo podrá establecerse una vez corregido el examen de diciembre 2024, pero un ejemplo ilustrativo puede verse en la Fig 9. Una vez fijada con los datos de diciembre 2024, esta ecuación de conversión se mantendrá sin cambios.

Fig 9



COMPARACIONES ENTRE CORRECCIÓN CON IRT Y MÉTODO TRADICIONAL

a) No habrá diferencias prácticas cuando se compare el puntaje entre examinados que rindieron el mismo examen: una persona que hubiese ranqueado en cierta posición con la corrección tradicional, obtendrá la misma posición con el nuevo sistema.

b) El puntaje obtenido estará siempre en proporción directa a la cantidad de respuestas correctas. En comparaciones con su misma promoción, a igual número de respuestas correctas, igual puntaje.

c) Sin embargo, la cantidad de respuestas correctas no corresponderá a igual puntaje cuando se comparen exámenes diferentes, ya que la cantidad de respuestas correctas exigidas para cada puntaje estará corregida (hacia arriba o hacia abajo), para ajustar la diferencia en dificultad de ambos exámenes.

d) Con el sistema tradicional no podían sacarse conclusiones del tipo “el examinado que obtuvo 75 puntos en el examen B tiene más conocimientos que un examinado que obtuvo 70 puntos en el examen A”. Con el nuevo sistema esto sí podrá inferirse.

e) Con el sistema tradicional, una escuela no podía concluir que el mayor puntaje de una promoción signifique mejora en el conocimiento de esa promoción. Con IRT y el procedimiento descrito aquí, esta inferencia será válida.

f) Con el sistema anterior, no podía asegurarse que los conocimientos mínimos necesarios para aprobar el examen fuesen los mismos, sólo podía asegurarse un mismo número de respuestas correctas. Con el nuevo sistema se asegura que el nivel de conocimientos mínimos para aprobar corresponda siempre al mismo.

(1) Un comité de expertos, compuesto de 12 representantes nacionales del norte, centro y sur, con cargos de responsabilidad en los distintos entornos clínicos de desempeño de un recién egresado (programas de especialización, hospitales de baja complejidad, consultorios municipales y servicios de urgencia), determinaron en un ejercicio real el requerimiento mínimo de habilidad para las competencias requeridas, que resultó ser de 0.1 logits. Nivel de habilidad que se hizo corresponder con los 51 puntos legales de aprobación de EUNACOM.