Universidad de Costa Rica
Facultad de Educación
Instituto de Investigación en Educación

FACULTAD DE EDUCACIÓN
INSTITUTO DE INVESTIGACIÓN EN EDUCACIÓN
(INIE)

INFORME FINAL

CONSTRUCCIÓN Y VALIDACIÓN DEL BANCO DE ÍTEMS PARA LA PRUEBA EXAMEN DE CONOCIMIENTOS GENERALES EN MEDICINA (ECGM) EMPLEADA PARA EL INGRESO AL PROGRAMA DE POSGRADO EN ESPECIALIDADES MÉDICAS
No. 724-B5-A79

Lucrecia Alfaro Rojas
Natalia Salas Segreda
María Paula Villarreal Galera

2016

Índice general

Índice de anexos

Índice de tablas

I. Información general administrativa

Número de la actividad aprobada: No. 7234-B5-A79

Nombre del proyecto: Construcción y validación del banco de ítems para la prueba “Examen de conocimientos generales en Medicina” (ECGM) empleada para el ingreso al Programa de Posgrado en Especialidades Médicas

Unidad base de las investigadoras: Sistema de Estudios de Posgrado (SEP)

Unidad de adscripción: Instituto de Investigaciones en Educación (INIE)

Programa al que pertenece: Cambio, desarrollo y gestión de la Educación Superior

Nombre de las investigadoras	Carga asignada	Programa que asigna la carga
Lucrecia Alfaro Rojas, Coord.	½	SEP
Natalia Salas Segreda	½	SEP (1/4), PPEM (1/4)
María Paula Villarreal	1/8	Ad honorem

Vigencia del proyecto: de 01/10/2015 a 30/09/2016

Resumen:

Esta actividad de investigación se enfoca en la generación de lineamientos para el proceso de construcción y manejo del ECGM empleado en el concurso nacional de ingreso al Programa de Posgrado en Especialidades Médicas. Apunta a la conformación de un banco de ítems para una prueba estandarizada con evidencias de validez, fundamentadas en las exigencias técnicas y científicas que la medición y evaluación psicoeducativas brindan para su desarrollo y aplicación, así como para el análisis de sus resultados.

La conformación de un banco de ítems implica un trabajo investigativo (desde el planteamiento del constructo teórico que está en la base de la prueba) y sistemático, no solo para la inclusión en el mismo de una cantidad mínima de ítems con evidencias de validez y confiabilidad, sino para su administración a mediano y largo plazo (diseño y pilotaje de nuevos ítems, alimentación continua del acervo existente, actualización de la tabla de especificaciones -contenidos, porcentajes, bibliografías a nivel mundial, estudios epidemiológicos y otros aspectos de la realidad nacional).

Descriptores: *EDUCACIÓN EN SALUD, Posgrados, pruebas, excelencia académica.*

II. Antecedentes

II.1. Introducción y antecedentes del proceso investigativo

De octubre de 2013 a setiembre de 2014, un equipo técnico conformado por profesionales en estadística, psicología y evaluación educativa se dio a la tarea de analizar cualitativa y psicométricamente el Examen de Conocimientos Generales en Medicina (ECGM) empleado en el concurso nacional para ingresar al Programa de Posgrado en Especialidades Médicas (PPEM) mediante la actividad de investigación No. 723-B3-771: Análisis cualitativo y psicométrico del "Examen de conocimientos generales en medicina" (ECGM) empleado en el concurso nacional para ingresar al Programa de Posgrado en Especialidades Médicas (PPEM).

La presente actividad de investigación, iniciada en octubre de 2015 y finalizada en el mes de setiembre del año en curso, se ha abocado a continuar con el trabajo desarrollado por el equipo investigador anterior, retomando las principales recomendaciones desprendidas de su análisis, así como adaptándose a las necesidades que han surgido en el proceso.

Tanto las autoridades del SEP, del PPEM y, durante este último año, del Instituto de Investigaciones en Educación (INIE), han considerado prioritario continuar con el desarrollo de las labores de construcción del banco de ítems a partir de los productos de la investigación anterior, es decir, se posibilita la evolución de un proceso que es continuo y tiene, hoy por hoy, un impacto en los procesos formativos de las diferentes especialidades médicas.

Entre los productos heredados que posibilitan la continuidad del proceso se encuentra la tabla de especificaciones de la prueba, matriz que refleja el constructo teórico que está en la base del diseño de los ítems. Cabe señalar que la elaboración de esta tabla generó espacios de trabajo grupal (en subcomisiones), a partir de los cuales los docentes del PPEM se han ido acercando, cada vez más, a la temática más amplia y siempre vigente de la medición y la evaluación educativas. Además, respondiendo a los procesos educativos, la tabla de especificaciones del ECGM, al tratarse de una prueba estandarizada, es dinámica y cambiante, por lo que cada año debe ser revisada y depurada por las personas expertas en el área de la mano de las investigadoras a cargo de este proyecto.

II.2. Planteamiento del problema

Para dimensionar el grado de relevancia de esta prueba, no solo es determinante el hecho de que la Universidad de Costa Rica es la única institución a nivel nacional con la responsabilidad de impartir estudios de posgrado en el Área de las Especialidades Médicas, sino la función particular del ECGM en el proceso de selección de las y los médicos generales que cursarán una especialización, es decir, de los futuros especialistas que tendrán a su cargo los servicios del Sistema Nacional de Salud de nuestro país.

En este sentido, la actividad actual se ha planteado resolver, con un enfoque investigativo y sistemático, el proceso de diseño, ensamblaje y aplicación del ECGM 2015-2016, en el marco de la colaboración UCR-CENDEISSS; por otro lado, se ha dado a la tarea de avanzar en el establecimiento de criterios y protocolos claros para el desarrollo de una prueba

estandarizada de altas consecuencias, de modo que estos aseguren la calidad y la confidencialidad en los procesos de construcción, validación y actualización de un banco de ítems.

Finalmente, desde una perspectiva investigativa más amplia, esta actividad no solo ha asumido un problema que amerita el diseño y la administración de una prueba bajo condiciones estandarizadas, sino que enfrenta el reto de propiciar una revisión permanente de los supuestos sobre la evaluación y la educación en Medicina que le subyacen.

III.3. Objetivos general y específicos

Nuestro objetivo general se planteó como la validación de ítems para la construcción de un banco de reactivos que permita el ensamblaje de futuras versiones de la prueba ECGM empleado en la primera etapa del concurso nacional para el ingreso al Programa de Posgrado en Especialidades Médicas.

Los objetivos específicos fueron los siguientes:

1- Sistematizar los protocolos de construcción y aplicación de la prueba “Examen de Conocimientos Generales en Medicina”

2- Capacitar constructores y jueces de los ítems para la prueba ECGM sobre aspectos relevantes para la construcción y análisis cualitativo de los mismo.

3- Brindar apoyo logístico para la aplicación de la prueba ECGM en 2016, siguiendo los criterios de estandarización que garanticen su calidad técnica.

4- Analizar las propiedades psicométricas de los ítems construidos e incluidos en la prueba ECGM aplicada en 2016.

5- Confeccionar el banco de ítems validados para la prueba ECGM.

En relación con el último objetivo, más que un banco final, esta actividad sentó las bases de un banco que debe seguir alimentándose.

A través de los objetivos expuestos, se busca que el sustento teórico y la calidad en la construcción de los ítems, así como la transparencia y la confidencialidad en su manejo, posibiliten un proceso de mejoramiento continuo del ECGM, tal como se propuso desde los primeros análisis. En este sentido, solo el trabajo conjunto entre el equipo técnico de esta prueba, el personal administrativo del PPEM y, primordialmente, el equipo experto de médicos comprometidos con la prueba y con el Posgrado hará posible la gestión de un banco de ítems que cumpla con los más altos estándares de calidad, a nivel nacional e internacional.

III. Referentes teóricos

Una prueba es un dispositivo evaluativo o un procedimiento mediante el cual se obtiene una muestra de la conducta de las personas examinadas en un dominio específico, el cual es posteriormente evaluado y calificado mediante un proceso estandarizado (AERA, APA y NCME, 2014, p. 2).

De acuerdo con la clasificación de Martínez, Hernández y Hernández (2006), se tiene el "Examen de conocimientos generales en Medicina" (ECGM) se clasifica como una prueba de altas consecuencias para las personas examinadas, ya que sus resultados definen el ingreso de estas al Posgrado en Especialidades Médicas de la Universidad de Costa Rica, lo que tiene importantes implicaciones en sus respectivos proyectos de vida.

Siguiendo con la clasificación, esta prueba también se describe como "de papel y lápiz" o de respuesta seleccionada, ya que la solución de los ítems consiste en elegir una opción dentro de un conjunto predeterminado de opciones de respuesta (una sola correcta y tres distractores, en este caso). Es una prueba de aplicación colectiva ya que evalúa simultáneamente una gran cantidad de personas; y es de potencia, pues se aplica en un contexto en el que el tiempo debe ser suficiente para contestar de manera reflexiva todos los ítems que, cabe señalar, tienen diferentes niveles de dificultad.

Finalmente, se ha considerado una prueba referida a normas en tanto las puntuaciones obtenidas por las personas examinadas se utilizan como medida de comparación entre todas ellas. Actualmente, las personas seleccionadas son aquellas con notas que se encuentran en y por encima del percentil 70 del total de examinados, por lo que no se puede definir la nota de aprobación sino con referencia al comportamiento de las notas del total de personas examinadas en un año específico. No obstante, es importante señalar que esta es una prueba construida con referencia a un conjunto de habilidades y conocimientos de los que deben dar cuenta las y los médicos generales que quieran ingresar al PPEM, es decir, existe un constructo teórico que tiene implicaciones directas en el diseño de la prueba, en la posición de la persona evaluada en relación con dicho constructo (no únicamente en comparación con las demás personas), así como en la interpretación psicométrica de los resultados obtenidos.

Esto último está en consonancia con las primeras actividades de investigación ligadas al ECGM, asimismo, propone un reto hacia el futuro en cuanto al modelo mixto de normas y criterios implicado en lo anteriormente expuesto.

El constructo teórico

De acuerdo con Leyva (2011), la definición de un constructo teórico (también llamado dominio o universo de medida en el caso de pruebas referidas a criterio) es origen y referencia de todas las etapas posteriores, por lo cual debe reunir características tales que permitan saber si un ítem (o reactivo) pertenece o no al dominio en cuestión. Por lo anterior, este constructo es el fundamento sobre el que se sostiene el diseño y desarrollo de una prueba a gran escala.

En términos generales, añade esta autora, existe coincidencia entre las personas expertas a la hora de afirmar que la calidad en la definición del dominio es lo que permite referir las

puntuaciones individuales de los ítems a criterios de estructura interna de la prueba. Esto a su vez es crucial a la hora de determinar si existen evidencias de validez de contenido y de constructo, así como otros conceptos de validez que se sugieren para este tipo de pruebas.

Para efectos del ECGM, se ha partido del postulado anterior en la forma de una tabla de especificaciones, tal y como la plantean Villarreal, Alfaro-Rojas y Brizuela (2015). Esta debe permitir a quienes construyen ítems conocer en detalle los temas y contenidos, habilidades cognitivas y tareas que se evaluarán con la prueba. Dicha tabla será además de gran utilidad para quienes deban diseñar y juzgar la calidad técnica de los ítems. En primer lugar, la tabla determinará si estos se ajustan o no al propósito de la prueba, así como a los contenidos, temas y subtemas planteados. En segundo lugar, pero no de menor importancia, la tabla de especificaciones servirá a las personas examinadas para conocer cuáles serán los contenidos evaluados, lo que les permitirá prepararse de manera más adecuada. Finalmente, la tabla será de la mayor relevancia para quienes empleen los resultados en la toma de decisiones, pues esta les permitirá fundamentar, a partir de un constructo teórico, las inferencias derivadas de los puntajes obtenidos.

Conceptos básicos sobre validez

Como se señaló, un concepto clave para la construcción e interpretación de resultados obtenidos mediante la aplicación de una prueba es el de validez. Esta se entiende como el grado en que la evidencia y la teoría respaldan las interpretaciones hechas a partir de los puntajes obtenidos por las personas examinadas en una prueba. La validez es, por lo tanto, una de las consideraciones más importantes en el desarrollo de una prueba y en lo que llamamos proceso de validación, el cual implica la acumulación de evidencias para proveer una base científica a dichas interpretaciones (AERA, APA y NCME, 2014).

En general, de acuerdo con Villarreal, Alfaro-Rojas y Brizuela (2015), hay dos aspectos indispensables a considerar: por un lado, la validez de las inferencias basadas en los puntajes obtenidos en una prueba; por otro, la justificación de los usos de esas inferencias. Estos aspectos no son compensatorios, es decir, lograr evidencias sobre uno de ellos no compensa las debilidades o carencias en cuanto al otro. Asimismo, afirman que la validez puede ser adecuada o no en la medida en que los puntajes obtenidos mediante la correcta aplicación de una prueba permitan fundamentar las inferencias en cuanto a lo que esta evalúa. De esta forma, el proceso de validación sería aquel mediante el cual se recaba, resume y evalúa la evidencia requerida para justificar las inferencias a partir de los resultados en una prueba.

Continuando con Villarreal, Alfaro-Rojas y Brizuela (2015), equipo investigador que forma parte de la presente actividad, la validación de las inferencias hechas a partir de una prueba requiere iniciar con una propuesta clara sobre las posibles interpretaciones que se pueden realizar a partir de esta, así como sus posibles usos. En este sentido, ambos deben basarse en un marco conceptual claro y explícito, ya que no es conveniente asumir que los potenciales usuarios de esta harán las interpretaciones correctas. Los autores señalan que lo anterior es de vital importancia si se toma en cuenta que la validez es una cuestión de grado y que puede cambiar con el tiempo en función de los avances teóricos, tecnológicos y metodológicos relacionados con el rasgo o característica psicológica (habilidad, actitud, conocimiento, patología, etc.) a medir, por lo que siempre es necesario actualizar la validez

de las interpretaciones y usos que se realizan con esta, así como incorporar la nueva evidencia que se vaya generando.

Ciertamente, recabar evidencias de validez requiere de un proceso continuo de investigación para fundamentar adecuadamente las inferencias realizadas con base en el uso de una prueba. Es importante señalar que las poblaciones se transforman constantemente en el transcurso de los años, por lo cual también es necesario actualizar continuamente los instrumentos de medición para adecuarse a los cambios que experimenten las poblaciones en la variable de interés y en otras características que puedan influir en la medición, tales como vocabulario, formato de los ítems, etc.

Actualmente se promueve una concepción unitaria de validez (Elosua, 2003; Montero, 2013), según la cual no existen distintos tipos de esta, sino diversas evidencias que se enfocan en diferentes facetas de aquella (AERA, APA y NCME, 2014). Esta evidencia puede ir referida a los contenidos de la prueba, su estructura interna, las estrategias de resolución de los ítems, la relación de la prueba con otras variables y las consecuencias del uso de esta, entre otros. A continuación, se mencionan algunas estrategias para recabar evidencias de validez según diferentes enfoques.

- Contenidos de la prueba

Para establecer la relevancia de los contenidos de una prueba respecto del constructo meta es fundamental, por una parte, haber hecho una labor investigativa profunda a la hora de diseñar la tabla de especificaciones en tanto matriz de ese constructo teórico que, a su vez, definirá la lógica y la proporcionalidad de los diferentes contenidos para el diseño de los ítems; por otra parte, se debe contar con el juicio de personas expertas que elaboren los ítems (en el caso de una prueba de conocimientos generales en Medicina) y posteriormente valoren si estos son una muestra adecuada de dichos contenidos en el área de interés (Martínez, Hernández y Hernández, 2006).

En relación con la tabla de especificaciones, cabe señalar que la labor investigativa no se acaba cuando esta queda establecida por primera vez. En un área como la Medicina donde los conocimientos se actualizan de manera permanente, esta tabla debe revisarse no solo con base en la última bibliografía pertinente, sino en los aspectos epidemiológicos y sociales de la realidad nacional. Esta revisión permanente es lo que permite, en el tiempo, seguir recabando evidencias sobre la pertinencia de los contenidos de la prueba y las inferencias realizadas a partir de los ítems.

- Estructura interna de la prueba

La identificación de los posibles patrones de asociación entre las respuestas a los ítems incluye métodos de análisis como el Análisis Factorial Exploratorio. Asimismo, existen métodos especializados para el análisis de la estructura interna de una prueba, la cual suele estar compuesta por ítems cuya distribución no es normal ni continua (Tate, 2003). A partir de las correlaciones entre las respuestas de todas las personas, estos métodos permiten identificar conjuntos homogéneos de ítems que deberían conformarse de acuerdo con los planteamientos teóricos que motivaron la creación de la prueba (Villarreal, Alfaro-Rojas y Brizuela, 2015), básicamente, de acuerdo con su tabla de especificaciones y las habilidades preestablecidas.

- Relación de la prueba con otras variables

Como se ha dicho, uno de los aspectos fundamentales respecto del tema de la validez es el grado en el que las puntuaciones de una prueba reflejan el constructo que se desea evaluar. Para indagar sobre este aspecto, Cronbach y Meehl (1955) plantearon que para establecer con claridad cuál variable se mide mediante una prueba, es necesario enmarcar dicha variable en una red de relaciones teóricas con otras variables, es decir, establecer un modelo a partir de esta relación y cuantificar su ajuste a los datos observados (Bollen y Hoyle, 2012). De acuerdo con Villarreal, Alfaro-Rojas y Brizuela (2015), existen técnicas para poner a prueba diferentes hipótesis sobre las relaciones entre una prueba y otras variables, como Modelos de Regresión, Modelos de Ecuaciones Estructurales, Análisis Factorial Exploratorio, Análisis Factorial Confirmatorio, entre otros. Según estos autores, otras estrategias que han sido utilizadas para proporcionar evidencias de validez (discriminante y convergente) son las Matrices Multirrasgo-Multimétodo y los modelos de regresión múltiple.

Consecuencias del uso de la prueba

Desde esta perspectiva, el desarrollo de una prueba exige recabar evidencias sobre la presencia de varianza irrelevante, así como los posibles efectos de la sub-representación de la variable de interés en el desempeño mostrado por las personas examinadas (Villarreal, Alfaro-Rojas y Brizuela, 2015). En este sentido, identificar posibles sesgos en contra de grupos o poblaciones específicos es parte de los controles de calidad de una prueba estandarizada.

El sesgo en contra de ciertas poblaciones es causado por todos aquellos componentes irrelevantes que resultan en menores puntajes para ciertos subgrupos de personas examinadas (AERA, APA y NCME, 2014), por lo que es necesario identificar las posibles diferencias en cuanto a la capacidad predictiva de una prueba entre grupos sociodemográficos, tasas diferenciales de selección de examinados con características irrelevantes al objetivo de una prueba, entre otros (Villarreal, Alfaro-Rojas y Brizuela, 2015).

Aunado a lo anterior, es importante determinar si personas con el mismo nivel en el rasgo evaluado presentan diferentes probabilidades de contestar correctamente los ítems que componen la prueba. Para ello, existen técnicas como el Análisis del Funcionamiento Diferencial del Ítem, el cual se realizó para efectos de esta actividad de investigación.

Conceptos básicos sobre confiabilidad

La confiabilidad de los puntajes obtenidos en una prueba se refiere al grado en que estos están libres del error de medición. Así, un instrumento es considerado como confiable si arroja resultados similares cuando es aplicado en diferentes momentos a un mismo conjunto de individuos examinados (Kumar, 2009). En otras palabras, la confiabilidad de un instrumento se ve reducida en la medida en que las proyecciones realizadas con este se ven afectadas por errores aleatorios de medición debidos a diversas circunstancias de los candidatos (cansancio, motivación, etc.) y del ambiente (temperatura, ruido, etc.) en el que responden los ítems (Ross y Rowley, 1991).

Autores como Martínez (1996) y Kane (2013) señalan que un alto nivel de consistencia en un instrumento para evaluar a las personas examinadas no garantiza la validez de las interpretaciones que se realizan con base en dicho instrumento. Villarreal, Alfaro-Rojas y Brizuela (2015) señalan que, en el ámbito de la medición psicoeducativa, la confiabilidad es una condición necesaria pero no suficiente para concluir que las inferencias a partir de una prueba son válidas.

Etapas en el diseño y la aplicación de una prueba estandarizada de altas consecuencias

Villarreal, Alfaro-Rojas y Brizuela (2015) sugieren implementar una serie de lineamientos generales para cada una de las etapas que componen el desarrollo de pruebas estandarizadas, particularmente las de selección única. Este equipo investigador determinó dichos lineamientos no solo con base en los estándares internacionales de AERA, APA, NCME (2014) y el *Educational Testing Service* (2002), sino considerando el estado de la cuestión de la medición educativa costarricense en lo relativo a las pruebas estandarizadas.

En tanto estos autores también forman parte del equipo a cargo de la presente actividad de investigación, se ha procedido a adaptar las etapas planteadas en su publicación al proceso particular del ECGM, de la siguiente manera:

1. Planificación de la prueba y aclaración (revisión) del propósito de la misma
2. Actualización de la tabla de especificaciones
3. Construcción y juzgamiento de ítems
4. Ensamblaje y embalaje de la prueba
5. Aplicación de la prueba bajo condiciones estandarizadas
6. Calificación de la prueba y manejo de las apelaciones
7. Análisis e investigación para recopilar evidencias de validez y confiabilidad
8. Alimentación y mantenimiento del banco de ítems
9. Generación y/o adaptación de manuales técnicos o protocolos

Las etapas señaladas se desprenden, por lo tanto, del marco teórico del que se ha dado cuenta en este apartado y, al mismo tiempo, han permitido generar la ruta metodológica, tal y como se detalla a continuación.

IV. Procedimientos metodológicos

En la medida en que esta actividad se abocó a desarrollar una prueba de altas consecuencias, el ECGM - convocatoria 2015-2016, se procuró seguir las etapas según el marco teórico propuesto, observando, de la manera más rigurosa posible, las recomendaciones de la aplicación del año anterior y, muy especialmente, los ajustes necesarios que el proceso requiriera.

IV. 1. Planificación, setiembre 2015

Un aspecto fundamental de esta etapa fue la incorporación del aporte del proyecto de investigación anterior, tal y como se ha venido planteando a lo largo de este informe.

En este sentido, se partió del propósito explícito y oficial de la prueba definido durante dicha investigación, a saber, "seleccionar a los médicos generales con el mayor grado de conocimientos[1] en Medicina General que deseen ingresar a alguno de los Programas de este Posgrado". Asimismo, se sistematizó la información dejada por el equipo anterior, con lo cual se empezó a enriquecer una base de datos digital con los ítems desarrollados durante el proceso 2014-2015 con un nuevo código que fuera compatible con los ítems por construir en el nuevo período.

Se actualizó un informe para la Decanatura del Sistema de Estudios de Posgrado con una síntesis de los resultados de dicha actividad, así como de lo esperado por esta nueva propuesta, lo anterior en el marco de la continuidad del proceso de aplicación del ECGM y en relación con los planes de estudio de las respectivas Especialidades Médicas (ANEXO 1. Presentación para la Decanatura del SEP).

Un aspecto innovador en esta convocatoria tiene que ver con su objetivo primordial de generar un banco de reactivos. En consecuencia, se ha decidido implementar la inclusión en la prueba de una cantidad de ítems experimentales, de modo que se puedan obtener datos psicométricos de los mismos y puedan incluirse o excluirse del banco a partir de evidencias científicas de validez y confiabilidad. Esto tendrá implicaciones en el ensamblaje de la prueba, en el proceso de calificación y, evidentemente, en la gestión del banco a mediano y largo plazo.

Cabe señalar que el modelo de medida con el que se analizará los datos obtenidos es la Teoría Clásica de los Tests (TCT).

IV. 2. Actualización de la tabla de especificaciones, setiembre-octubre 2015

Al igual que el propósito de la prueba, la tabla de especificaciones se estableció durante la actividad de investigación anterior, de acuerdo con las siguientes etapas: a- identificación de áreas del conocimiento en Medicina General y habilidades cognitivas a evaluar; b- 15 sesiones con expertos de análisis por área de conocimiento para definir contenidos: c- diseño

1 Como mayor grado de conocimientos se entiende "la capacidad para evaluar, diagnosticar, tratar o referir a partir de conocimientos en Medicina General", según lo establecido como el propósito de esta prueba por la Comisión Coordinadora del PPEM.

y construcción de la tabla de especificaciones con sus respectivos temas y contenidos, así como sus pesos relativos y la bibliografía a emplear como referente teórico para la construcción de los ítems.

En setiembre-octubre 2015, se actualizó dicha tabla con base en una discusión de la bibliografía empleada en la convocatoria anterior, así como en la revisión de las proporciones de los diferentes temas y subtemas planteados. De este modo, la tabla quedó actualizada por área después de las sesiones (ANEXO 2. Tabla de especificaciones por área, actualización 2015-2016).

IV. 3. Construcción y juzgamiento de ítems, octubre-noviembre-diciembre 2015

Como se señaló en la introducción de este informe, desde la actividad anterior se estableció una metodología de trabajo por subcomisiones (una por cada área de conocimiento del ECGM), las cuales han sido de suma importancia en los procesos de construcción y juzgamiento de los reactivos que hoy alimentan el banco de ítems.

Una vez actualizada la tabla de especificaciones por área, con la cantidad de ítems a construir por temas y subtemas, y estos distribuidos en la proporción establecida (Cirugía, Medicina Interna, Ginecología-Obstetricia y Pediatría, cada una conformando un 20% de la prueba, y Psiquiatría y Medicina Familiar y Comunitaria, cada una un 10%), se procedió a la capacitación de las y los médicos constructores de ítems. Se llevaron a cabo 6 reuniones con este propósito: 4 en octubre, 1 en noviembre y 1 en diciembre. Se generó una presentación con algunos criterios para la construcción de ítems, así como ejemplos concretos, los cuales se adaptaron según la Especialidad a la que estaba dirigida la capacitación (Ver ANEXO 3. Algunos criterios para la construcción de ítems para el ECGM).

Durante el mes de octubre y la primera quincena de noviembre, se abrió el periodo de recepción de los ítems construidos por parte de las y los doctores expertos en los diferentes temas del ECGM. Estos se entregaron en una boleta facilitada para tales efectos (ver ANEXO 4. Boleta para la construcción de ítems), la información fue digitalizada posteriormente por el Equipo Técnico de la prueba. Las personas constructoras debieron firmar un documento donde se comprometían a destruir cualquier borrador previo a la entrega del ítem, pues este es propiedad de la Universidad de Costa Rica (ver ANEXO 5. Compromiso de confidencialidad).

Durante los meses de noviembre y diciembre se llevaron a cabo dos reuniones por área (12 en total), denominadas juzgamientos. Estos se realizaron con el propósito de revisar, discutir y, de ser necesario, corregir los ítems construidos durante las semanas previas. Para estos efectos, se trabajó con dos formatos: uno para el equipo técnico a cargo de la sesión (ver ANEXO 6. Boleta de ítem para juzgamiento, Equipo Técnico), donde se puede apreciar la versión original del ítem, una versión con cambios propuestos por parte del Equipo Técnico, la respuesta correcta y la justificación, tanto de la opción de respuesta correcta como de las incorrectas; el segundo formato se le entregaba a los jueces expertos para su resolución, con la versión con cambios, sin la indicación de la respuesta correcta y con un espacio para la valoración del ítem como “muy difícil”, “difícil”, “dificultad media”, “fácil” o “muy fácil” (ver ANEXO 7. Boleta de ítem para juzgamiento por Área, Equipo Médico).

Finalmente, en diciembre de 2015, se llevó a cabo un juzgamiento final, con todas las áreas presentes (al menos el o la Coordinadora de cada Área), con el propósito de valorar un examen completo, no solo desde la perspectiva de sus propios ítems, sino de la de las demás Áreas, asegurando con esto que el nivel de la prueba fuera, efectivamente, el de conocimientos generales en Medicina.

IV. 4. Ensamblaje y embalaje de la prueba, enero-febrero 2016

En esta convocatoria 2015-2016, cabe resaltar la novedad del pilotaje de ítems para dar inicio a la creación del banco, según el objetivo primordial de esta actividad de investigación.

Esto quiere decir también que, en esta oportunidad, aún no se contaba con datos psicométricos de ninguno de los ítems disponibles, por lo que el ensamblaje de la prueba se hizo con base en las evaluaciones por parte de los médicos expertos y del nivel de dificultad de los ítems durante los juzgamientos por Área. En este sentido, se procuró contar con un balance entre ítems muy difíciles y fáciles, así como un grueso de ítems de dificultad media.

Se trabajó con 4 fórmulas diferentes, cada una de ellas con 120 ítems calificados y 30 experimentales. Los ítems calificados fueron idénticos en las 4 fórmulas y se agruparon por Área. En cada fórmula, podía variar el orden de las Áreas, mas no así el orden de los ítems dentro de una misma Área. Los 30 ítems de carácter experimental fueron diferentes en cada fórmula, lo que permitió pilotear un total de 120 ítems (con un cuarto de la población total evaluada), ordenados por Área, de manera que este pilotaje fuera también un reflejo de la tabla de especificaciones. Estos ítems fueron otorgados como correctos a todas las personas aspirantes. De este modo, no solo quedan fuera del proceso posterior de apelaciones (con lo que se mantienen en estricta confidencialidad), sino que se actúa de manera consecuente con el principio de que solo aquellos ítems con buenos resultados psicométricos pasan a formar parte del banco.

En cuanto al embalaje y otros aspectos de logística de la prueba, el Equipo Técnico trabajó de la mano con el personal administrativo del PPEM, tanto en la configuración de las bases de datos para la posterior calificación en el Sistema de Ingreso al PPEM (SIPPEM), como en la impresión y preparación del material con la totalidad de las pruebas e instrucciones del caso.

IV. 5. Aplicación de la prueba bajo condiciones estandarizadas, marzo 2016

Como señalan Villarreal, Alfaro-Rojas y Brizuela (2015), una prueba se considera estandarizada cuando se construye y administra respetando normas preestablecidas y bajo condiciones sistematizadas y equiparables.

Desde la convocatoria anterior, la prueba ECGM se aplica en las instalaciones de la Universidad de Costa Rica, con condiciones de aplicación similares en todas las sedes. Esto propicia no solo un entorno académico, como conviene a una prueba de estas características, sino una serie de condiciones ambientales óptimas para la concentración de las personas evaluadas.

Asimismo, se capacitó a un grupo de personas coordinadoras de sede (según los edificios universitarios previamente reservados) y de personas aplicadoras por equipos, cada uno a cargo de un número de aulas y de estudiantes predeterminados, de acuerdo con el padrón de las personas inscritas en la prueba. Cada equipo fue entrenado para organizar y mantener el orden en los diferentes espacios a su cargo.

Una vez aplicada la prueba, el manejo de las cajas (instrucciones, folletos y hojas de respuesta) se llevó a cabo mediante los más estrictos controles de seguridad y confidencialidad (como para otras etapas de este proceso, puede observarse el manual de procedimientos en el ANEXO 8 de este informe).

IV. 6. Calificación de la prueba y manejo de las apelaciones, marzo-abril 2016

La calificación se llevó a cabo a partir del archivo de claves con las respuestas manejada por el Equipo Técnico de la prueba. Posteriormente, esta información se trasladó a las bases de datos del personal administrativo del PPEM y del Centro de Informática de la Universidad de Costa Rica, con quienes se ha venido trabajando para manejar los resultados de las personas aspirantes. Para esta convocatoria, por primera vez se trabajó con el sistema para la presentación y resolución de apelaciones en línea, lo que simplificó el trabajo de los constructores de ítems apelados. Anteriormente esto se resolvía en formato físico y de manera manual, lo que presentaba grandes desventajas: primero, en cuanto al volumen de apelaciones (cajas de apelaciones en los más diversos formatos); segundo, a nivel del riesgo para la seguridad misma de ese material (resguardo y confidencialidad de una gran cantidad de documentos que debían ser llevados a las diferentes oficinas de los constructores responsables); finalmente, en cuanto se impedía el anonimato en el proceso (el constructor podía conocer la identidad de las personas que apelaban, con lo cual el sesgo en la decisión sobre la apelación se acrecienta considerablemente).

Por lo anterior, las consecuencias del manejo de las apelaciones por medio del SIPPEM es uno de los cambios más significativos en la convocatoria 2015-2016.

IV. 7. Análisis e investigación para recopilar evidencias de validez y confiabilidad, mayo-junio 2016

Una vez aplicada la prueba se elaboró una base de datos con las respuestas dadas a los ítems por parte de las personas examinadas. A partir de esta base, se llevaron a cabo análisis factoriales exploratorios, así como la estimación de los coeficientes de consistencia interna de la prueba y los índices de discriminación y dificultad para cada uno de los ítems mediante la Teoría Clásica de los Tests (TCT), todo esto mediante el programa estadístico SPSS. También, se realizó el análisis para la detección del Funcionamiento Diferencial de los Ítems (FDI) mediante el paquete estadístico STATA.

Los resultados de estos análisis se recopilan en el apartado V. Análisis y discusión de los resultados.

IV. 8. Alimentación y mantenimiento del banco de ítems

Según Villarreal, Alfaro-Rojas y Brizuela (2015), el banco de ítems es un archivo (físico o digital) en el que se almacenan los ítems que cumplen con todos los requerimientos psicométricos para evaluar los contenidos o el constructo de interés.

Actualmente, el PPEM cuenta con un archivo digital con la información de cada ítem: la versión original del ítem y la versión con los cambios finalmente avalados después de su evaluación en los diferentes comités de juzgamiento (por área y final). Por otra parte, la información sobre las propiedades psicométricas de cada ítem se encuentra actualmente en cédulas individuales, tanto en formato físico como digital. Esto con el propósito de permitir a futuros desarrolladores encontrar de una manera fácil y rápida las características de los ítems disponibles, lo cual deberá agilizar el diseño y el ensamblaje de las futuras versiones de la prueba. No se presentará en este informe el texto de los ítems que conforman el banco, debido a que este es un material de carácter confidencial, que se utilizará para el ensamblaje de pruebas de altas consecuencias en un futuro próximo. No obstante, sí se describe el comportamiento general de los ítems y se adjunta una tabla con la información de los que conformaron el ECGM 2016 (ANEXO 9. Tabla con información general sobre los ítems aplicados en el ECGM 2016).

A continuación, se presentan algunas consideraciones básicas sobre el banco de ítems actual:

1. Los ítems están archivados en ampos físicos adecuadamente resguardados, así como en una base de datos digital, en un disco duro también resguardado. Desde las últimas semanas, se encuentran también en un sistema informático diseñado para ese fin (denominado SIPPEM, Sistema de Ingreso al Programa de Posgrado en Especialidades Médicas), con lo cual, se tiene la oportunidad de actualizar los ítems directamente desde el sistema en línea. Esto permite reimprimirlos y mantenerlos actualizados en cualquiera de sus formatos y con sus respectivas estadísticas.

2. El archivo de ítems está organizado por áreas y temas, de acuerdo con la tabla de especificaciones de la prueba y por año de construcción.

3. En el banco físico (ampo), cada ítem corresponde a una hoja. En esta aparece el ítem completo (versión original y versión con cambios), la respuesta correcta señalada y la bibliografía. En una cédula adjunta se puede ver sus datos estadísticos (dificultad, discriminación, análisis de distractores, apelaciones totales, apelaciones rechazadas y apelaciones concedidas).

4. Dentro del conjunto de ítems que componen el banco, se ha señalado el año de aplicación. Con esto se evita que un mismo ítem se exponga de nuevo demasiado pronto a la población meta.

5. Proponemos incrementar el banco anualmente con los ítems experimentales que presenten propiedades psicométricas adecuadas. Actualmente contamos con 58 ítems que fueron experimentales y que, por lo tanto, podrán ser utilizados en las próximas convocatorias, según se determine en el momento del ensamblaje.

6. De los ítems probados en el ECGM 2016 se rechazaron 61 (36 calificadas y 25 experimentales) debido a que tuvieron una discriminación baja (incluso, negativa) o a que todas las apelaciones relativas a cada uno de ellos fueron concedidas, lo que los define como ítems de poca calidad. Por otra parte, 62 ítems se clasificaron en la categoría “re-experimentar” debido a que presentaron problemas con los distractores o porque, mediante las apelaciones presentadas, se detectó aspectos a mejorar en el encabezado o en las opciones de respuesta, sin que esto los invalidara del todo. Esto deja como resultado 117 ítems con suficiente calidad técnica para ser utilizados en ensamblajes futuros del ECGM.

7. El lugar donde se ubica actualmente el banco físico de ítems es seguro (se cierra mediante llave) y está acondicionado para el almacenamiento de documentos, de manera que solamente tienen acceso a este los miembros del equipo técnico de la prueba. El manejo de los ítems mediante el sistema informático (SIPPEM) se realiza con todas las garantías de seguridad informática y solamente las personas del equipo técnico y administrativo del PPEM pueden accederlo, mediante claves personalizadas.

IV. 9. Generación y adaptación de los manuales técnicos

Para este apartado, véase el ANEXO 8. Manual de procedimientos, Examen de Conocimientos Generales en Medicina, ECGM, 1era etapa.

Es fundamental señalar que los procesos 2015-2016 se llevaron a cabo manualmente, por lo que los protocolos respectivos responden a ese procedimiento. A partir de junio de 2016 se ha venido desarrollando el sistema de construcción en línea como parte de la plataforma del SIPPEM, lo que, inevitablemente, implicará reformas en estos manuales.

V. Análisis y discusión de los resultados

V.1. Análisis descriptivo de la población que realizó el ECGM 2016

Se presenta a continuación un resumen de las características socio-demográficas de los profesionales que realizaron el ECGM en el año 2016, como examen de primera etapa en el concurso de ingreso a una especialidad médica en el 2017.

En total, 1.805 médicos realizaron el examen en el año 2016. Al inicio del proceso, se registraron 2.186 aspirantes, de los cuales 254 no se inscribieron a la prueba ya que habían pasado a la segunda etapa en algún concurso anterior. De las restantes 1.932 personas inscritas, no se presentaron a aplicar la prueba 127.

Del total de 1.805 médicos a los que se aplicó la prueba, poco más de la mitad correspondió a mujeres, concretamente un 56,07%, mientras que un 43,93% correspondió a hombres.

De las personas que hicieron la prueba, la mayoría fueron costarricenses, únicamente el 2,22% estuvo conformado por extranjeros.

La distribución de personas evaluadas según universidad (Tabla 1) muestra aquellas universidades de las cuales son egresados las y los médicos evaluados. Una cuarta parte de

estas personas egresaron de la UCIMED; los centros de educación superior que le siguen, con mayor número de aspirantes en orden decreciente, son la UNIBE, la Universidad Latina y la UCR. De esta última proviene solamente el 11,47% de los médicos evaluados en esta prueba.

Tabla 1. Distribución de personas evaluadas según universidad, ECGM 2016

Universidad	Absoluto	Relativo (%)
UCIMED	443	24,54
UNIBE	327	18,12
ULatina	266	14,74
UCR	207	11,47
UH	176	9,75
UIA	162	8,98
UACA	127	7,04
Extranjera	54	2,99
SJT	42	2,33
UAC	1	0,06
Total	1805	

El año de graduación de las personas evaluadas va de 1993 a 2016, pero la mayoría (55,96%) se graduó entre 2013 y 2016.

El número de especialidades a las cuales desearon ingresar los médicos inscritos fue en total 44. De estas, las que despertaron mayor interés fueron Ginecología y Obstetricia (10,97%), Anestesiología y Recuperación (8,25 %), Radiología e Imágenes Médicas (8,03%), Cirugía General (6, 98%) y Pediatría (6.15%).

Los datos que se presentan en los siguientes subapartados corresponden exclusivamente a los resultados obtenidos con los 1.805 profesionales que realizaron la prueba en el año 2016.

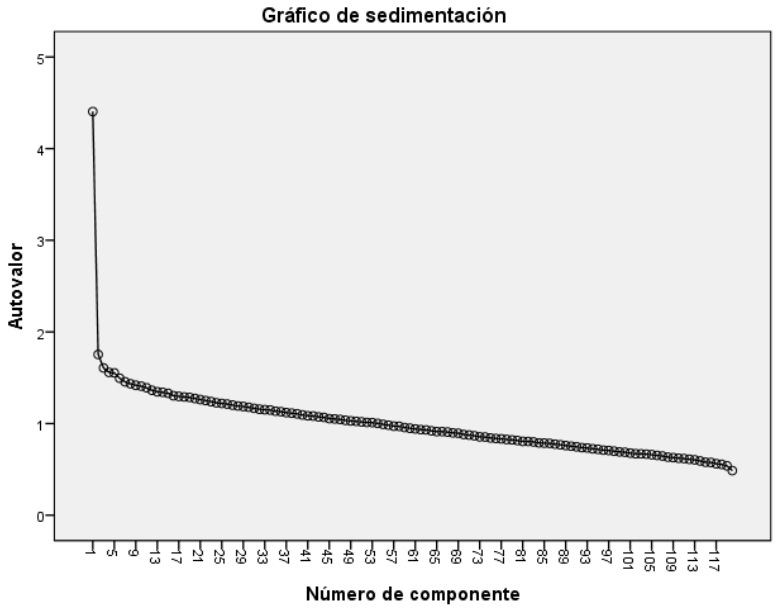
V.2. Análisis psicométrico del ECGM 2016

Con base en los resultados de los 1.805 aspirantes evaluados en el año 2016, se procedió a realizar un análisis psicométrico de los ítems bajo la Teoría Clásica de los Tests (TCT). A continuación se presentan diferentes análisis, a saber: factorial exploratorio, de confiabilidad, de discriminación y dificultad de los ítems, finalmente, de distractores y funcionamiento diferencial de los ítems.

V.2.1. Análisis factorial exploratorio (AFE)

Con el fin de encontrar evidencias que apunten a la unidimensionalidad en el constructo que se mide, se efectuó un análisis factorial exploratorio (AFE). Se realizó el análisis considerando solamente las preguntas que se repitieron en las 4 fórmulas (120 ítems), mediante la técnica de análisis por componentes principales y utilizando la rotación Promax. La medida de adecuación muestral de Kaiser-Meyer-Olkin fue de 0,669. Y, por medio de la prueba de esfericidad de Bartlett, se obtuvo χ^2 = 11795,441 con 7.140 gl (p<0,001). El porcentaje de variancia explicado del primer factor es de un 3,67%. Por lo tanto, no hay evidencia estadística suficiente para afirmar que existe al menos un factor común entre los ítems. El gráfico de sedimentación es otro criterio considerado para determinar el número de factores y, en este, se puede ver que no se cumple con el supuesto (Gráfico 1).

Gráfico 1. Gráfico de sedimentación del análisis de factores de los ítems del ECGM 2016



V.2.2. Análisis de confiabilidad (coeficiente alfa de Cronbach)

Se calculó el valor del coeficiente del alfa de Cronbach para los 120 ítems comunes a las 4 fórmulas. Para este análisis no se consideraron los ítems experimentales, ya que estos no son comunes a la totalidad de las fórmulas aplicadas. Se obtuvo que el valor alcanzado fue de 0,68, el cual es un valor inferior al esperado para una prueba de altas consecuencias como el ECGM (0,8). Este hallazgo se interpreta como un llamado a tomar medidas para aumentar la consistencia interna de la prueba.

A pesar de este resultado inicial, es posible aumentar el coeficiente de confiabilidad de la prueba si se eliminan del análisis todos aquellos ítems que tuvieron una baja discriminación (ver el análisis de discriminación en el apartado siguiente). En el caso del ECGM 2016, se detectó que 38 de los 120 ítems no aportaban a la consistencia interna de la prueba. Así, al

eliminarlos del análisis, se logró aumentar el coeficiente de confiabilidad a 0,75, lo que lo acerca a la medida meta para este coeficiente.

Como recomendación, se espera que, en el futuro, no se tomen en cuenta para la calificación todos aquellos ítems que tengan baja discriminación, de manera que se pueda contar con el mejor coeficiente de confiabilidad posible para la prueba.

V.2.3. Análisis de discriminación y dificultad de los ítems

Después de una valoración en detalle de cada uno de los ítems que conformaron las 4 fórmulas del ECGM 2016, se obtuvo el índice de dificultad y el índice de discriminación para cada uno de los 240 utilizados. En el caso de los 120 ítems calificados, el índice de dificultad se calculó con base en las respuestas de los 1.805 aspirantes que respondieron a las fórmulas, mientras que, en el caso de los ítems experimentales, el cálculo se basó en la totalidad de personas que contestaron la fórmula en la que aparecía el ítem en cuestión.

En cuanto a la dificultad del ECGM 2016, se encontró que casi la mitad de los ítems (el 48,33%) se consideró fácil o muy fácil y que hubo poca representatividad de ítems con dificultad media (Tabla 2). La categoría de ítems difíciles sí correspondió a la cantidad recomendada de ítems de este tipo en una prueba estandarizada.

Tabla 2. Clasificación de ítems del ECGM 2014, de acuerdo con su índice de dificultad

Nivel de dificultad	Ítems calificados	%	Ítems experimentales	%	Total ítems	%
Muy fácil	26	21,67	28	23,33	54	22,50
Fácil	30	25,00	32	26,67	62	25,83
Dif. Media	13	10,83	15	12,50	28	11,67
Difícil	26	21,67	25	20,83	51	21,25
Muy difícil	25	20,83	20	16,67	45	18,75

En cuanto al índice de discriminación, este se calculó para los 240 ítems. Al igual que para la dificultad, se hizo el análisis en dos partes: una referida a los 120 ítems comunes a las 4 fórmulas y otra referida a los ítems experimentales exclusivos a cada una de ellas. Después de obtener el indicador de discriminación para cada ítem, esta se clasificó de acuerdo con la codificación sugerida: Muy buena (≥ 0,40), Buena (0,30 a 0,39), Regular (0,20 a 0,29) y Deficiente (≤ 0,19). Según la tabla 3, es evidente que ningún ítem obtuvo el indicador de "Muy buena" discriminación, solamente 4 de ellos se pueden considerar con "Buena"

discriminación y 43 con un nivel “Regular” de discriminación. De los ítems con un índice de discriminación “Deficiente”, se detectaron 193, 29 de los cuales tuvieron un índice de discriminación negativo, es decir, favorecieron a personas con bajo desempeño general en la prueba.

Tabla 3. Clasificación de ítems del ECGM-2016 de acuerdo con su índice de discriminación

Nivel de discriminación	Ítems calificados	%	Ítems experimentales	%	Total Ítems	%
Muy buena	0	0	0	0	0	0
Buena	0	0,00	4	3,33	4	3,33
Regular	19	15,83	24	20,00	43	35,83
Deficiente	101	84,17	92	76,67	193	160,83

Aunque no existe un porcentaje ideal de ítems en cada una de las categorías sugeridas, es evidente que debe mejorarse la capacidad para discriminar entre personas con niveles de habilidad bajos y altos, ya que uno de los objetivos de la prueba es primero excluir los ítems de baja calidad, para después seleccionar a los aspirantes que pasarán a la segunda etapa del proceso de ingreso al PPEM.

V.2.4. Análisis de distractores

Para cada ítem se analizó si alguno presentó distractores que fueran seleccionados por las personas evaluadas casi en la misma o en mayor proporción que la opción correcta. Se encontró que, en total, hubo 48 ítems en los que se presentó esta situación. De estos casos, 12 fueron preguntas calificadas (no experimentales), mientras que 36 fueron experimentales. El que la respuesta correcta no sea la más seleccionada da señales de problemas de construcción de los ítems, por lo que se remarca la importancia de experimentar con ellos antes de que sean utilizados como calificados en una prueba. De esta forma se garantiza su calidad.

V.2.5. Funcionamiento Diferencial del ítem (DIF)

El análisis de impacto permite observar si existen diferencias reales entre hombres y mujeres en relación con el constructo que se mide con esta prueba; este debe realizarse de manera previa a verificar si existe funcionamiento diferencial del ítem (o DIF, por sus siglas en inglés).

En este caso, el puntaje de la prueba presentó diferencias significativas entre el nivel de habilidad por género. Se realizó una comparación de medias en la que se encontró que

existen diferencias estadísticamente significativas entre hombres y mujeres (t(1803)=-2,319, p=0,02). Los intervalos de confianza, o los límites probables entre los que se encuentra la verdadera diferencia entre las dos medias de ambos puntajes, se muestran como IC 95% [0,141; 1,683].

Lo anterior permite decir que la verdadera diferencia entre la nota promedio de los hombres y la nota promedio de las mujeres se encuentra entre 0,141 y 1,683. Este dato (además de que el intervalo no incluye 0) permite rechazar la hipótesis sobre igualdad de medias. No obstante, se estima el tamaño del efecto (la *d* de Cohen) en 0,110, lo que, en la literatura estadística, es considerado como un valor muy pequeño. Esto implica que tampoco existen diferencias grandes entre estas dos medias.

Tabla 4. Medias y tamaños del efecto para la Nota en el ECGM 2016

Sexo	N	Media	D.E
Femenino	1012	60.1057	8.325
Masculino	793	61.0177	8.247

Planteado lo anterior, se procede a realizar el presente estudio bajo la premisa de la equidad en los análisis psicométricos del ECGM. Mediante el paquete estadístico Winsteps, se detecta el funcionamiento diferencial de los ítems (DIF) con el método empírico Mantel-Haenszel (MH) y con el método propio del Modelo de Rasch (MR). Las investigadoras de este proyecto consideraron importante identificar aquellos ítems en los que, mediante ambos métodos, MH y MR, se detectara la presencia de DIF (Tabla 5).

Los resultados muestran que, de los 120 ítems que contempló la prueba, 35 presentaron DIF (29.16%). Esto es: las personas que los contestaron tienen distinta probabilidad de acertar el ítem según su género, aun cuando estas tienen un mismo nivel de habilidad en el constructo medido. Considerando en detalle los 35 ítems del cuadro anterior, 18 resultaron con DIF a favor de los hombres y 17 a favor de las mujeres. Aunque esto no significa que deba bajarse la alerta ante los ítems con DIF, es importante destacar que, en este caso, se afectó a ambas poblaciones de manera similar.

Si bien con base en los resultados de este análisis no se tomaron decisiones inmediatas sobre si desechar o no ítems, la información generada se tomará como base para futuros análisis cualitativos que permitan analizar aquellos factores presentes en los ítems que puedan marcar alguna tendencia a la inequidad en el ECGM.

Tabla 5. Análisis del Funcionamiento Diferencial de los Ítems (DIF) del ECGM 2016

ÍTEM	Θ Mujer	S.E. Mujer	θ Hombre	S.E. Hombre	Contraste	S.E.	Prob. Rasch	Prob. Mantel Hanszel	Tamaño Mantel Hanszel	A favor de cuál sexo
Ítem2	-0,09	0,06	-0,35	0,07	0,26	0,10	0,0090	0,0121	0,26	Hombres
Ítem5	-1,86	0,09	-1,50	0,09	-0,36	0,13	0,0063	0,0101	-0,31	Mujeres
Ítem10	2,08	0,10	1,76	0,10	0,32	0,14	0,0210	0,0162	0,33	Hombres
Ítem11	0,94	0,07	1,19	0,08	-0,25	0,11	0,0236	0,0239	-0,26	Mujeres
Ítem15	-0,69	0,07	-0,38	0,07	-0,31	0,10	0,0018	0,0020	-0,31	Mujeres
Ítem19	0,34	0,06	0,13	0,07	0,21	0,10	0,0324	0,0219	0,23	Hombres
Ítem20	0,57	0,07	0,83	0,08	-0,26	0,10	0,0093	0,0328	-0,18	Mujeres
Ítem24	-0,33	0,06	-0,57	0,08	0,24	0,10	0,0172	0,0260	0,2	Hombres
Ítem27	0,07	0,06	-0,25	0,07	0,32	0,10	0,0012	0,0009	0,32	Hombres
Ítem29	-0,36	0,07	-0,59	0,08	0,23	0,10	0,0216	0,0374	0,19	Hombres
Ítem39	0,52	0,07	0,30	0,07	0,22	0,10	0,0235	0,0207	0,22	Hombres
Ítem40	0,53	0,07	0,01	0,07	0,52	0,10	0,0000	0,0000	0,51	Hombres
Ítem41	1,25	0,08	0,91	0,08	0,34	0,11	0,0019	0,0020	0,3	Hombres
Ítem46	2,39	0,11	1,78	0,10	0,61	0,15	0,0000	0,0000	0,57	Hombres
Ítem52	0,19	0,06	-0,06	0,07	0,25	0,10	0,0095	0,0085	0,23	Hombres
Ítem53	-0,46	0,07	-0,14	0,07	-0,32	0,10	0,0012	0,0015	-0,29	Mujeres
Ítem58	0,76	0,07	1,00	0,08	-0,24	0,10	0,0232	0,0435	-0,21	Mujeres
Ítem61	-0,46	0,07	-0,07	0,07	-0,39	0,10	0,0001	0,0001	-0,37	Mujeres
Ítem62	0,30	0,06	0,77	0,08	-0,47	0,10	0,0000	0,0000	-0,45	Mujeres
Ítem71	0,98	0,07	1,46	0,09	-0,48	0,11	0,0000	0,0000	-0,46	Mujeres
Ítem80	-2,03	0,10	-1,60	0,10	-0,43	0,14	0,0018	0,0010	-0,46	Mujeres
Ítem88	-2,45	0,12	-2,05	0,11	-0,40	0,16	0,0124	0,0051	-0,5	Mujeres
Ítem90	-1,37	0,08	-0,76	0,08	-0,61	0,11	0,0000	0,0000	-0,73	Mujeres
Ítem91	0,37	0,06	0,03	0,07	0,34	0,10	0,0005	0,0010	0,32	Hombres
Ítem99	-1,86	0,09	-1,30	0,09	-0,56	0,13	0,0000	0,0000	-0,55	Mujeres
Ítem102	-1,52	0,08	-1,09	0,08	-0,43	0,12	0,0003	0,0000	-0,53	Mujeres
Ítem106	-2,19	0,10	-1,70	0,10	-0,49	0,14	0,0008	0,0001	-0,57	Mujeres
Ítem110	-0,96	0,07	-1,19	0,09	0,23	0,11	0,0391	0,0463	0,23	Hombres
Ítem120	1,78	0,09	1,22	0,08	0,56	0,12	0,0000	0,0000	0,54	Hombres
Ítem121	0,18	0,06	-0,04	0,07	0,22	0,10	0,0234	0,0330	0,23	Hombres
Ítem127	1,08	0,07	0,72	0,08	0,36	0,10	0,0006	0,0006	0,35	Hombres
Ítem134	0,42	0,07	0,01	0,07	0,41	0,10	0,0000	0,0000	0,43	Hombres
Ítem136	2,27	0,10	2,69	0,14	-0,42	0,17	0,0156	0,0185	-0,36	Mujeres
Ítem146	1,14	0,07	0,87	0,08	0,27	0,11	0,0109	0,0096	0,27	Hombres
Ítem150	0,30	0,06	0,62	0,07	-0,32	0,10	0,0011	0,0005	-0,37	Mujeres

θ: Habilidad medida siguiendo el modelo de Rasch.

VI. Conclusiones y recomendaciones

El trabajo realizado durante esta actividad de investigación hizo posible el ensamblaje y la aplicación exitosa del ECGM en el año 2016, de manera que se pudo evaluar los conocimientos generales en medicina de 1.805 profesionales provenientes de las diferentes escuelas de Medicina del país. Todos estas personas aspiraban a ingresar a alguna de las especialidades impartidas en el PPEM. La calidad y la organización del proceso hicieron posible una meticulosa recolección de la información en condiciones que garantizaron la estandarización del proceso. Lo anterior respalda la generación de datos y los análisis que dan cuenta de una adecuada medición del constructo meta.

Por su parte, el cálculo del indice de discriminación permitió detectar aquellos ítems que no aportaron a la consistencia interna del instrumento. La exclusión de estos ítems en la calificación hubiera permitido elevar el valor del coeficiente de confiabilidad (alfa de Cronbach) del conjunto de ítems comunes a las 4 fórmulas de 0,68 a 0,75, siendo este último un valor más aceptable para una prueba de altas consecuencias como el ECGM. Ante la necesidad de mejorar la capacidad para discriminar entre personas con niveles de habilidad bajos y altos, con el propósito de seleccionar a aquellas que pasarán a la segunda etapa del proceso de ingreso al PPEM, se recomienda que, en aplicaciones futuras, se excluya los ítems con baja discriminación en una etapa previa a la calificación final, de manera que el instrumento tenga más poder para seleccionar a aquellas personas con mayor puntaje en relación con el constructo medido.

El análisis de dificultad del instrumento arrojó que, si bien casi la mitad de los ítems de la prueba aplicada en 2016 pueden considerarse como fáciles, sí se contó con una adecuada representación de ítems con dificultad alta. La poca representatividad de ítems con dificultad media es un reto a tomar en cuenta en futuros procesos de construcción de ítems y de ensamblaje de la prueba, de manera que se promueva más variabilidad de la dificultad en los diferentes ítems del instrumento.

Dado que, en esta ocasión, el análisis factorial exploratorio no arrojó evidencias que apunten a la existencia de, al menos, un factor común entre los ítems, se hace necesario profundizar en la discusión a lo interno del equipo investigador, acerca de la pertinencia de esperar unidimiensionalidad en una prueba que mide un constructo tan amplio y con tantas facetas, como es el caso de los conocimientos generales en medicina. Paralelamente a esto, se debe continuar haciendo esfuerzos por mejorar la calidad de los ítems y clasificarlos de acuerdo con tareas más concretas (“Valoración del riesgo e Interpretación diagnóstica”, “Abordaje terapéutico” y “Seguimiento y evolución”), de manera que se pueda valorar el agrupamiento por factores que correspondan a estas categorías.

Por su parte, el análisis de distractores permite explorar posibles problemas en la construcción de ítems, por lo que se subraya la importancia de seguir incluyendo, en futuras aplicaciones, ítems con carácter experimental para revisar y corregir problemas de construcción antes de que se conviertan en preguntas calificadas en una prueba.

El análisis de impacto permitió explorar la existencia de diferencias entre el desempeño del grupo de hombres y mujeres evaluados con el ECGM. La comparación de medias arrojó que existen diferencias estadísticamente significativas entre ambos sexos, sin que la diferencia

de medias pueda considerarse grande y sin que se haya favorecido solo a uno de estos grupos. En efecto, se detectaron 35 ítems que presentaron DIF, de los cuales 18 resultaron favorables para los hombres y 17 para las mujeres. A partir de estos datos, se puede sostener que ambas poblaciones se vieron afectadas de manera similar. Ciertamente, la información generada se tomará como base de futuros análisis cualitativos, con el propósito de combatir aquellos factores presentes en los ítems que puedan tender a la inequidad en el ECGM. Esto solo podrá hacerse mediante un análisis del contenido de los ítems que presentaron DIF, ya que dicho contenido constituye la principal fuente de información cualitativa (más allá de los indicadores acerca de su acierto por parte de las personas que los respondieron) para valorar las diferencias en el desempaño obtenido por ambos grupos.

Como parte del trabajo realizado, se sistematizaron los protocolos de construcción y aplicación del ECGM, de manera que ya se cuenta con un documento depurado que orienta los procesos relativos a la construcción y ensamblaje anual de la prueba. Cabe señalar que, con el advenimiento del sistema en línea (SIPPEM), dichos manuales tendrán que ajustarse.

En relación con el proceso de construcción de la prueba, las coordinaciones realizadas con los diferentes programas de especialidad del PPEM permitieron capacitar a una cantidad adecuada de constructores de ítems, los cuales, posteriormente, aportaron reactivos para el instrumento. Esto permitió mejorar la calidad de los ítems construidos y el juzgamiento de los mismo por parte de las diferentes comisiones. En relación con la aplicación de la prueba, se brindó apoyo logístico al equipo humano que la llevó a cabo, siguiendo los criterios de estandarización que dan calidad del instrumento.

El proceso de construcción, análisis y almacenamiento de los ítems construidos, muchos de los cuales fueron incluidos en el ECGM 2016, permite afirmar que ya se está en el proceso de consolidación de un banco de ítems, como señala uno de los principales objetivos de esta actividad de investigación. Ciertamente, se ha logrado implementar un proceso amplio y ordenado que, con el seguimiento adecuado, llevará a la consolidación de un banco de ítems para el ECGM. Este banco deberá crecer con el paso del tiempo mediante la incorporación de nuevos ítems, su experimentación previa y el descarte de aquellos que vayan perdiendo su capacidad de discriminar entre las personas con los conocimientos generales en Medicina y mejores candidatas para el ingreso a las Especialidades Médicas.

A modo de conclusión general, es posible afirmar que los alcances de este proyecto benefician, en primera instancia, a las personas graduadas en Medicina y Cirugía General, aspirantes a cursar los programas de posgrado en Especialidades Médicas, esto mediante un proceso de selección con un alto grado de calidad técnica, a partir de criterios de estandarización y equidad. Por otra parte, es fundamental considerar que también hay un beneficio para las y los médicos de la CCSS, docentes del PPEM: no solo se ha implementado un sistema de resolución más expedita y justa de las apelaciones, sino que el proceso como tal les ha provisto de conocimientos sobre medición y evaluación que, a su vez, podrán extender tanto al diseño de las pruebas de segunda etapa como a sus prácticas docentes. Finalmente, está claro que habrá un beneficio colateral para las personas usuarias de los servicios de la CCSS, al contar con procesos de calidad que aseguren la selección óptima de los médicos especialistas en cuyas manos colocaremos la responsabilidad de velar por la salud de un país.