UNIVERSIDAD NACIONAL AUTÓNOMA DE MÉXICO
PROGRAMA DE MAESTRÍA Y DOCTORADO EN CIENCIAS MÉDICAS, ODONTOLÓGICAS Y DE LA SALUD
FACULTAD DE MEDICINA
CAMPO DE CONOCIMIENTO: CIENCIAS SOCIOMÉDICAS
EDUCACIÓN EN CIENCIAS DE LA SALUD

Evaluación de los reactivos de opción múltiple utilizados en los exámenes sumativos de la asignatura de Inmunología en la Facultad de Medicina de la UNAM

TESIS
QUE PARA OPTAR POR EL GRADO DE
MAESTRO EN CIENCIAS

PRESENTA:

Jesús Rivera Jiménez

TUTOR O TUTORES PRINCIPALES

Dr. Adrián Alejandro Martínez González
Facultad de Medicina

Dr. Fernando Flores Hernández
Facultad de Medicina

Ciudad Universitaria, Cd. Mx. octubre de 2018

Dedicatorias

A mis padres, porque gracias a ellos soy y estoy.

A todos los estudiantes, quienes me han motivado a participar en mejorar la calidad educativa en mi Facultad.

A la UNAM, por brindarme un segundo hogar y un espacio para mi formación y crecimiento académico y personal.

Agradecimientos

En primer lugar a mis padres, Clara Jiménez y Jesús Rivera, quienes siempre han estado ahí para apoyarme y que, gracias a ellos, he podido llegar hasta este momento de mi vida. Sin ustedes, nada de esto hubiera sido posible.

Gracias al Dr. Raúl Chávez, quien desde el primer momento de mi incorporación como instructor de la asignatura de Inmunología depositó su confianza en mí y que gracias a él he podido crecer académicamente, abriéndome las puertas de la Coordinación de Inmunología y apoyando cada una de mis decisiones dentro del Departamento. Agradezco también al Dr. Edgar Zenteno y al Dr. Juan Pablo Pardo, Jefes del Departamento, quienes también apoyaron este proyecto y todos los cambios derivados de él para mejorar el proceso educativo en la asignatura de Inmunología.

Gracias al Dr. Adrián Martínez, tutor principal de mi proyecto, y al Dr. Melchor Sánchez, Coordinador del posgrado en Educación en Ciencias de la Salud, por su confianza desde el inicio de este proyecto y por inspirarme para poder crecer y trascender en el camino de la educación médica.

Gracias a todos los estudiantes que me han inspirado y me han permitido ser parte de su formación. El esfuerzo derivado de este trabajo ha sido siempre en aras de mejorar la calidad educativa en la Facultad de Medicina. Gracias en particular a todos los médicos pasantes que me han apoyado en este camino y que a lo largo del tiempo han estado y siguen estando a mi lado: Arely Reyes y Gerardo Patiño (mis Topos); Gerardo Peña, Óscar Alencaster y Fausto Gómez (mis Pollos); Antonioni Ortega y Miguel Cuéllar (mis Cucos); Carlos Chávez, José Luis Maldonado y Denisse Jiménez (mis Bebos); Isaac Vásquez y Cynthia Fortozo (mis Bojos); Eduardo Santana, Rafael González y Alejandra Pando (mis Pandos); y Silvana Castelán (mi Chaka); todos ellos entrañables amigos, grandes médicos, así como compañeros incondicionales y piezas fundamentales en el crecimiento de la Inmunología en la Facultad.

Gracias a Oli Espinosa por su confianza y amistad en todo este tiempo. Eres una inspiración y modelo a seguir para mí. Mi mayor admiración.

Gracias a Iván Umbral, Fernando Rivera, Daniel Chávez y Adrián Cruz porque el tiempo a su lado me ha permitido madurar y aprender de mis errores. Gracias por su apoyo durante toda mi formación como médico y como educador.

Gracias a los amigos y familia que han estado a mi lado, que me han apoyado, respetado y aconsejado. Un agradecimiento especial a mi hermano, Raúl Torres, por siempre apoyarme y creer en mí.

Resumen

Introducción. La evaluación del aprendizaje es un aspecto fundamental del proceso educativo que, de acuerdo con los modelos actuales, permite valorar el logro de las competencias de los estudiantes. Ante la necesidad de realizar una evaluación válida, confiable y objetiva en la Facultad de Medicina de la UNAM, resulta necesario realizar un análisis de los instrumentos de evaluación que se utilizan con ese fin.

Método. Se realizó un estudio descriptivo y correlacional, transversal y retrospectivo de las características cualitativas y cuantitativas de los reactivos de opción múltiple utilizados en los exámenes sumativos de la asignatura de Inmunología en la Facultad de Medicina de la UNAM. Se desarrolló un instrumento para la evaluación de las características cualitativas y se obtuvieron las evidencias de validez relacionadas con su aplicación. Se utilizó la teoría clásica de los tests para realizar el análisis psicométrico como parte de las características cuantitativas de los reactivos.

Resultados. La clasificación de los reactivos se realizó de acuerdo con las características evaluadas, identificando entre 31 y 40% de reactivos que necesitan modificarse para poder ser parte del banco de reactivos de la asignatura.Al comparar los valores psicométricos de los reactivos estándar (sin errores en su elaboración) contra los reactivos con errores, se identificó que existen diferencias en la dificultad y en la discriminación de ambos tipos.

Conclusión. Estas diferencias en los parámetros psicométricos de los reactivos asociadas a errores en su elaboración pueden modificar la decisión de aprobación de un examen tanto a favor de un estudiante que no posee el aprendizaje, como en contra de uno que sí lo adquirió, lo que corresponde a una amenaza importante a la validez del proceso de evaluación.

Abstract

Introduction. Assessment of learning is a fundamental aspect along the educational process which, according to current models, allows to determinate the achievement of the competencies acquired by the students. Given the need for a valid, reliable and objective assessment in the Faculty of Medicine of the National Autonomous University of Mexico (UNAM), it is necessary to perform an analysis of the instruments used for this purpose.

Method. This is a descriptive and correlational, cross-sectional and retrospective study of qualitative and quantitative characteristics of multiple-choice items used in the summative examinations of the Immunology course in the Faculty of Medicine of UNAM. An instrument was designed for the assessment of quality features, and evidence of validity related to its application was gathered. The classical theory of the tests was performed for the psychometric analysis within the quantitative features.

Results. Items classification was conducted according to the characteristics assessed, identifying between 31 and 40% of items that needs to be modified in order to be part of the item bank. Comparing psychometric values of items with flaws against items without them, it showed differences in the difficulty and discrimination of both types.

Conclusion. These differences among the psychometrical parameters on the test items may modify the decision of passing one test in favor of a student who does not have the knowledge, or not passing the test against one student that it does acquired it, which corresponds to a major threat to the validity of the assessment process.

Contenido

INTRODUCCIÓN 1
1. MARCO TEÓRICO 3
1.1 EVALUACIÓN EDUCATIVA 3
1.2 EVALUACIÓN DEL APRENDIZAJE 5
1.3 INSTRUMENTOS DE EVALUACIÓN 7
1.4 REACTIVOS DE OPCIÓN MÚLTIPLE 10
1.5 CARACTERÍSTICAS CUALITATIVAS DE LOS REACTIVOS DE OPCIÓN MÚLTIPLE 12
1.5.1 VALIDEZ 12
1.5.2 CONSTRUCTO 14
1.5.3 TABLA DE ESPECIFICACIONES 15
1.5.4 RECOMENDACIONES PARA LA ELABORACIÓN DE REACTIVOS DE OPCIÓN MÚLTIPLE 15
1.6 CARACTERÍSTICAS CUANTITATIVAS DE LOS REACTIVOS DE OPCIÓN MÚLTIPLE 19
1.6.1 TEORÍAS DE LOS TESTS 19
1.6.2 ANÁLISIS DE ÍTEMS 20
2. ANTECEDENTES 23
2.1 LA FACULTAD DE MEDICINA Y EL PLAN DE ESTUDIOS 2010 23
2.2 LA ASIGNATURA DE INMUNOLOGÍA EN LA LICENCIATURA DE MÉDICO CIRUJANO DE LA FACULTAD DE MEDICINA DE LA UNAM 24
2.3 EVALUACIÓN DE LA ASIGNATURA DE INMUNOLOGÍA 25
3. MÉTODO 28
3.1 PLANTEAMIENTO DEL PROBLEMA 28
3.2 OBJETIVOS 29
3.3 JUSTIFICACIÓN 30
3.4 DISEÑO DE INVESTIGACIÓN 32
3.5 SELECCIÓN DE LA MUESTRA 32
3.6 RECOLECCIÓN DE DATOS 34
3.7 EVALUACIÓN CUALITATIVA DE LOS REACTIVOS 35
3.7.1 EVIDENCIA DE VALIDEZ RELACIONADA CON EL CONTENIDO 35
3.7.2 EVIDENCIA DE VALIDEZ RELACIONADA CON EL PROCESO DE RESPUESTA 37
3.7.3 EVIDENCIA DE VALIDEZ RELACIONADA CON LA ESTRUCTURA INTERNA 38
3.8 EVALUACIÓN CUANTITATIVA DE LOS REACTIVOS 39
3.9 ASPECTOS ÉTICOS 42
4. RESULTADOS 43
4.1 EVALUACIÓN CUALITATIVA DE LOS REACTIVOS 43

4.1.1 Evidencia de validez relacionada con el contenido. 43
4.1.2 Evidencia de validez relacionada con el proceso de respuesta. 43
4.1.3 Evidencia de validez relacionada con la estructura interna 45
4.1.4 Características cualitativas de los reactivos 46
4.2 Evaluación cuantitativa de los reactivos **52**
4.2.1 Parámetros psicométricos 52
4.2.2 Distractores no funcionales 57
4.3 Características cualitativas y cuantitativas de los reactivos **58**
4.3.1 Clasificación de reactivos de acuerdo con sus características 58
4.3.2 Calificación del estudiante de acuerdo con el tipo de reactivo 58
5. DISCUSIÓN **64**
5.1 Consideraciones sobre el instrumento para la evaluación de reactivos[70] **64**
5.2 Características cualitativas de los reactivos **68**
5.3 Características cuantitativas de los reactivos **70**
5.4 Evaluación de los reactivos de opción múltiple **74**
5.5 Limitaciones del estudio **76**
6. CONCLUSIONES **78**
7. REFERENCIAS **79**
8. ANEXOS **87**
Anexo 1. Recomendaciones para la elaboración de reactivos de opción múltiple. Haladyna, Downing, Rodríguez. A review of multiple-choice item-writing guidelines (2002) (traducción de CENEVAL)[55] **87**
Anexo 2. Tabla de especificaciones de la asignatura de Inmunología para el año escolar 2014 ..... **88**
Anexo 3. Comparación de recomendaciones para elaboración de reactivos **96**
Anexo 4. Primera propuesta de instrumento para evaluar reactivos de opción múltiple con 24 indicadores **101**
Anexo 5. Interfaz del cuestionario para obtener evidencia de validez por el juicio de expertos para el diseño del instrumento **102**
Anexo 6. Instrumento preliminar para la evaluación de reactivos de opción múltiple **103**
Anexo 7. Programa académico del Curso-Taller "Elaboración de reactivos para la evaluación del aprendizaje de Inmunología" **104**
Anexo 8. Distribución de número de respuestas correctas contestando ciegamente para 60 ítems con 4 opciones de respuesta **107**
Anexo 9. Distribución de número de respuestas correctas contestando ciegamente para 70 ítems con 4 opciones de respuesta **108**

ANEXO 10. CORRELACIONES ENTRE LAS PAREJAS DE ÍTEMS DEL INSTRUMENTO PARA EVALUAR REACTIVOS DE OPCIÓN MÚLTIPLE 109
ANEXO 11. ANÁLISIS DE CONFIABILIDAD DEL INSTRUMENTO PRELIMINAR..... 110
ANEXO 12. PRUEBA T PARA LOS CRITERIOS DEL INSTRUMENTO PRELIMINAR 111
ANEXO 13. ANÁLISIS FACTORIAL DEL INSTRUMENTO PARA EVALUAR REACTIVOS DE OPCIÓN MÚLTIPLE. 113
ANEXO 14. CLASIFICACIÓN DE REACTIVOS DE ACUERDO CON DOWNING (2009) Y HALADYNA (2013) ... 115
ANEXO 15. CORRELACIONES ENTRE EL NÚMERO DE DISTRACTORES NO FUNCIONALES (NFD), DIFICULTAD Y DISCRIMINACIÓN DE LOS REACTIVOS 118
ANEXO 16. PRUEBA DE LOS RANGOS CON SIGNO DE WILCOXON PARA COMPARAR LA DIFICULTAD (DIF) Y DISCRIMINACIÓN (DISC) ENTRE REACTIVOS ESTÁNDAR (ST) Y REACTIVOS CON ERRORES (FL)..... 119
ANEXO 17. CORRELACIÓN ENTRE NÚMERO DE ERRORES Y PARÁMETROS PSICOMÉTRICOS 120
ANEXO 18. EJEMPLO DE LA BASE DE DATOS DE LAS RESPUESTAS DE LOS ESTUDIANTES EN LOS EXÁMENES EVALUADOS. 121

Índice de tablas

PÁGINA

TABLA 1.1 DIFERENTES DEFINICIONES DE LOS NIVELES TAXONÓMICOS DEL DOMINIO COGNITIVO DE BLOOM 18

TABLA 3.1 DISTRIBUCIÓN DE REACTIVOS, DURACIÓN DE LA ASIGNATURA Y NÚMERO DE EXÁMENES APLICADOS EN LOS PERIODOS A ANALIZAR 34

TABLA 3.2 CLASIFICACIÓN DE REACTIVOS DE ACUERDO CON SU DIFICULTAD Y DISCRIMINACIÓN (TRADUCIDO DE DOWNING & YUDKOWSKY, 2009) 39

TABLA 3.3 CLASIFICACIÓN DE REACTIVOS DE ACUERDO CON SU DIFICULTAD Y DISCRIMINACIÓN (TRADUCIDO DE HALADYNA, 2013) 40

TABLA 4.1. CONCORDANCIA INTERJUECES EN LA PROPUESTA DE INSTRUMENTO 44

TABLA 4.2 CORRELACIONES ENTRE CRITERIOS DEL INSTRUMENTO 45

TABLA 4.3 MATRIZ DE ESTRUCTURA DE ANÁLISIS DE COMPONENTES PRINCIPALES PARA EL INSTRUMENTO FINAL PARA LA EVALUACIÓN DE REACTIVOS DE OPCIÓN MÚLTIPLE 47

TABLA 4.4. COMPORTAMIENTO PSICOMÉTRICO DE LOS REACTIVOS 54

TABLA 4.5 FRECUENCIA DE REACTIVOS DE ACUERDO CON LA CLASIFICACIÓN DE DOWNING 57

TABLA 4.6 FRECUENCIA DE REACTIVOS DE ACUERDO CON LA CLASIFICACIÓN DE HALADYNA 57

TABLA 4.7 FRECUENCIA DE DISTRACTORES NO FUNCIONALES EN LOS 200 REACTIVOS ANALIZADOS 57

TABLA 4.8 CLASIFICACIÓN DE LOS 200 REACTIVOS DE ACUERDO CON LA PROPUESTA DE DOWNING (2009) Y EL NÚMERO DE ERRORES EN SU ELABORACIÓN 59

TABLA 4.9 CLASIFICACIÓN DE LOS 200 REACTIVOS DE ACUERDO CON LA PROPUESTA DE HALADYNA (2013) Y EL NÚMERO DE ERRORES EN SU ELABORACIÓN 59

TABLA 4.10 ESTADÍSTICOS DE LAS CALIFICACIONES OBTENIDAS EN LOS EXÁMENES, CLASIFICADAS DE ACUERDO CON LOS TIPOS DE REACTIVOS QUE CONFORMAN CADA TIPO DE PRUEBA 60

TABLA 4.11 CORRELACIONES PARA CADA PAREJA DE LOS TIPOS DE REACTIVOS QUE CONFORMAN LOS EXÁMENES ANALIZADOS 61

TABLA 4.12 ANÁLISIS DE REGRESIÓN LINEAL PARA ESTIMAR EL GRADO DE PREDICCIÓN DEL PUNTAJE GLOBAL DE CADA UNO DE LOS TIPOS DE REACTIVO QUE CONFORMAN EL EXAMEN 62

TABLA 4.13 COMPORTAMIENTO PSICOMÉTRICO DE LOS DIFERENTES TIPOS DE REACTIVOS EN EL PRIMER EXAMEN, CONFORMADO POR 60 REACTIVOS 62

TABLA 4.14 COMPORTAMIENTO PSICOMÉTRICO DE LOS DIFERENTES TIPOS DE REACTIVOS EN EL SEGUNDO EXAMEN, CONFORMADO POR 70 REACTIVOS 62

TABLA 4.15 COMPORTAMIENTO PSICOMÉTRICO DE LOS DIFERENTES TIPOS DE REACTIVOS EN EL TERCER EXAMEN, CONFORMADO POR 70 REACTIVOS 62

TABLA 4.16 COMPORTAMIENTO PSICOMÉTRICO DE LOS DIFERENTES TIPOS DE REACTIVOS TOMANDO EN CUENTA EL TOTAL ANALIZADO (200 REACTIVOS) 63

TABLA 4.17 NÚMERO DE ESTUDIANTES APROBADOS CONSIDERANDO LA CALIFICACIÓN GLOBAL (GL) Y LA OBTENIDA CON LOS REACTIVOS ESTÁNDAR (ST) 63

Índice de figuras

	PÁGINA
FIGURA 1.1 PIRÁMIDE DE MILLER	10
FIGURA 4.1 DISTRIBUCIÓN DE LOS NIVELES TAXONÓMICOS ENTRE LOS 308 REACTIVOS EVALUADOS	48
FIGURA 4.2 HISTOGRAMA QUE MUESTRA LA DISTRIBUCIÓN DE REACTIVOS DE ACUERDO CON EL NÚMERO DE ERRORES EN SU ELABORACIÓN, A PARTIR DEL INSTRUMENTO PRELIMINAR DE 21 INDICADORES	49
FIGURA 4.3 HISTOGRAMA QUE MUESTRA LA DISTRIBUCIÓN DE REACTIVOS DE ACUERDO CON EL NÚMERO DE ERRORES EN SU ELABORACIÓN, A PARTIR DEL INSTRUMENTO FINAL DE 14 INDICADORES	50
FIGURA 4.4 PORCENTAJE DE APEGO A LOS CRITERIOS DEL INSTRUMENTO PRELIMINAR	50
FIGURA 4.5 PORCENTAJE DE APEGO A LOS CRITERIOS DEL INSTRUMENTO FINAL	51
FIGURA 4.6 DISTRIBUCIÓN DE REACTIVOS DE ACUERDO CON EL NÚMERO DE ERRORES EN SU ELABORACIÓN, CONSIDERANDO ÚNICAMENTE LOS 200 REACTIVOS SELECCIONADOS PARA EL ANÁLISIS PSICOMÉTRICO	52
FIGURA 4.7 FRECUENCIA DE ÍNDICES DE DIFICULTAD ENTRE LOS REACTIVOS EVALUADOS	54
FIGURA 4.8 FRECUENCIA DE ÍNDICES DE DISCRIMINACIÓN ENTRE LOS REACTIVOS EVALUADOS	55
FIGURA 4.9 ÍNDICES DE DIFICULTAD PARA CADA UNO DE LOS EXÁMENES ANALIZADOS, CONSIDERANDO LOS DIFERENTES TIPOS DE REACTIVOS (GL, ST Y FL)	55
FIGURA 4.10 ÍNDICES DE DISCRIMINACIÓN PARA CADA UNO DE LOS EXÁMENES ANALIZADOS, CONSIDERANDO LOS DIFERENTES TIPOS DE REACTIVOS (GL, ST Y FL)	55
FIGURA 4.11 CONFIABILIDAD DE CADA UNO DE LOS EXÁMENES ANALIZADOS, CONSIDERANDO LOS DIFERENTES TIPOS DE REACTIVOS (GL, ST Y FL)	56

Introducción

El Plan de Estudios 2010 de la licenciatura de Médico Cirujano de la Facultad de Medicina de la UNAM representa un reto educativo para la institución, ya que el cambio a un modelo educativo con un enfoque por competencias implica una reestructuración del ejercicio docente para lograr su adecuada implementación.

Dentro de los grandes retos en el proceso educativo se encuentra el mejorar los instrumentos utilizados para la evaluación del aprendizaje de los estudiantes. Los exámenes con reactivos de opción múltiple han sido uno de los instrumentos más utilizados a nivel internacional para la evaluación del dominio cognitivo de las competencias (que corresponde a los niveles en donde se evalúa al estudiante únicamente el "saber" del tema o en ocasiones el "saber cómo", de acuerdo con la Pirámide de Miller sobre el desarrollo de la competencia clínica)[1]. Si bien este recurso está limitado a la evaluación de los niveles más bajos de la competencia clínica, esto no le resta importancia ya que son el fundamento para el desarrollo de niveles superiores de la misma, por lo que es imperativo que exista suficiente evidencia de validez sobre el proceso de evaluación realizado con su uso.

Existe una vasta gama de referentes sobre recomendaciones para elaborar este tipo de reactivos, y existe también una gran evidencia de que no existe un apego a las mismas durante su elaboración, lo cual puede tener consecuencias negativas para el estudiante.

En este trabajo se describe el proceso que permitió llevar a cabo los siguientes puntos:

- Se analizan las características que tienen los reactivos utilizados en los exámenes parciales sumativos de la asignatura de Inmunología durante el año escolar 2014.

- Se propone un instrumento que permite evaluar las características de los reactivos de opción múltiple y se describen las evidencias de validez relacionadas con su uso en la evaluación de reactivos en el área básica de las ciencias de la salud.
- Se analiza el comportamiento psicométrico de los reactivos al ser aplicados en los estudiantes, tomando como referente la Teoría Clásica de los Test.
- Se identifican aquellos reactivos que cumplen tanto las características cualitativas y cuantitativas adecuadas para su incorporación al banco de reactivos de la asignatura, de acuerdo con las recomendaciones propuestas en la literatura.
- Se analiza la relación entre el uso de reactivos con errores en su elaboración y la calificación obtenida por el estudiante.

Antes de la descripción del proyecto, se plantean los principios teóricos sobre evaluación y medición pertinentes para el desarrollo del mismo, así como los antecedentes del entorno educativo en el que se llevó a cabo el estudio.

1. Marco teórico

1.1 Evaluación educativa

De acuerdo con el Diccionario de la Real Academia Española[2], la evaluación se define como "acción y efecto de evaluar", y al respecto, evaluar es definido como "estimar, apreciar, calcular el valor de algo" y también como "estimar los conocimientos, aptitudes y rendimiento de los alumnos". Este último concepto está enfocado directamente en el proceso educativo, ya que la acción se dirige hacia uno de los componentes del mismo. Scriven[3] define a la evaluación como el proceso de determinar el mérito y valor de las cosas, y la refiere como un proceso analítico clave en todas las disciplinas intelectuales y prácticas.

Según el documento *Standards for Educational and Psychological Testing*[4], la evaluación se define como "cualquier método sistemático de obtención de información utilizado para realizar inferencias sobre las características de las personas, objetos o programas". La evaluación es entonces un proceso intrínseco de toda la actividad educativa y tiene un campo muy amplio de acción que permite la toma de decisiones. Dentro de la evaluación educativa se pueden identificar diversas áreas de especialización, que incluyen la evaluación del aprendizaje, de la docencia, de los materiales educativos, de los programas educativos y de las instituciones educativas[5]. Existen tres términos en inglés que se refieren a estos diversos aspectos de la evaluación[6]; es conveniente hacer la aclaración, ya que en ese idioma cada término tiene un significado diferente. *Assessment* se refiere a la emisión de juicios del progreso individual de un estudiante y el logro de sus objetivos de aprendizaje. *Appraisal* se refiere a los juicios del desempeño de los profesionales de la educación, como profesores, líderes escolares, entre otros. También puede referirse a la pertinencia, factibilidad y sostenibilidad potencial de una intervención para el desarrollo antes de tomar la decisión de otorgar un financiamiento. *Evaluation*, como tal, se utiliza para referirse a los juicios, tan sistemáticos y objetivos como sean posibles, que se emiten sobre la efectividad de las escuelas,

las políticas, los sistemas y los programas educativos. Su objetivo es determinar la pertinencia y el logro de los objetivos, así como la eficiencia, la eficacia, el impacto y la sostenibilidad para el desarrollo.

De acuerdo con la Organización para la Cooperación y el Desarrollo Económicos (OCDE)[6], se está dando a nivel mundial una mayor importancia al proceso de evaluación dentro de las actividades educativas, lo que se ve reflejado en un crecimiento de la evaluación educativa en los sistemas escolares, ya que las políticas educativas dan una mayor importancia a estos temas, y por ello se han creado entidades dedicadas exclusivamente al control de los procesos de evaluación. Debido a esto, hay una mayor variedad de actividades de evaluación dentro de las instituciones educativas, ya que deja de ser la evaluación de los estudiantes el único objetivo, y se pone énfasis en la evaluación de los programas y de la institución, así como la valoración de los docentes. Ante esta situación, se está otorgando una mayor importancia al desarrollo de indicadores y a la medición educativa; esto implica que los resultados de los estudiantes son el referente para comparar la efectividad de los sistemas educativos a nivel internacional, los logros de los objetivos de acuerdo con la institución educativa y la relación del desempeño docente con el progreso de los estudiantes; esto implica también el desarrollo de un número mayor de indicadores para medir el desempeño de las escuelas, a partir de diversos aspectos demográficos, administrativos y contextuales. Otro aspecto fundamental en esta transición es que se está dando un mayor número de usos a los resultados obtenidos en los procesos de evaluación, ya que son una fuente importante de información acerca de los logros de aprendizaje de los estudiantes, y ofrece información también a los padres y a la sociedad sobre el rendimiento educativo y las mejoras en el desempeño de los docentes y las escuelas; esta creciente relevancia que surge del uso de los resultados de la evaluación implica una mayor responsabilidad de los participantes en el proceso de medición de los resultados obtenidos por los estudiantes, como son profesores, escuelas y legisladores. Con lo anterior, se pueden crear incentivos para aquellos desempeños sobresalientes e identificar deficiencias dentro de los sistemas escolares.

Otra cuestión a destacar de este mismo reporte de la OCDE es que actualmente existe una internacionalización de la evaluación, ya que los datos del desempeño de los estudiantes han influido de manera importante en las discusiones nacionales sobre educación, permitiendo la promoción de reformas en las políticas educativas de diversos países; el establecimiento de estándares internacionales sobre lo que los estudiantes deben saber (estándares de contenido) y lo que deben de ser capaces de hacer (estándares de desempeño) ha permitido que los diferentes países realicen adaptaciones propias a partir de estos referentes. La sofisticación tecnológica que existe en la actualidad ha permitido la realización de evaluaciones a gran escala, una mejor medición de habilidades cognitivas como la solución de problemas, con enfoques más personalizados en la evaluación, y una mayor confiabilidad y menores costos en la aplicación de exámenes a los estudiantes.

1.2 Evaluación del aprendizaje

La evaluación del aprendizaje (también conocida como evaluación del estudiante) es un campo especializado de la evaluación educativa. Se refiere a cualquier actividad en la que se obtienen evidencias de aprendizaje de una manera planeada y sistemática, lo cual se usa para emitir un juicio sobre el aprendizaje[7]. En este tipo de evaluación se valoran los conocimientos, habilidades y actitudes adquiridas por los estudiantes como resultado de diversas experiencias educativas[5]. Bernard[8] define a la evaluación como “la medida de los niveles de mejora que en el plano del conocimiento y de las habilidades cognitivas personales aparecen en la conducta de los estudiantes como consecuencia de las experiencias vividas en el aula y fundamentalmente de lo que hacen para alcanzar los objetivos educativos asignados a la institución escolar a través de la programación académica. En inglés, se utiliza el término *assessment* para referirse a este tipo de evaluación.

Existen diferentes clasificaciones de este tipo de evaluación, algunos son los siguientes[5,9]:

Tipos de evaluación de acuerdo al objetivo. Aquí se encuentran la *evaluación sumativa* y la *evaluación formativa*. La evaluación sumativa se encarga de recapitular el aprendizaje obtenido por el alumno, con el fin de certificar los logros alcanzados; suele aplicarse al final de una intervención o de una fase de esta misma; también proporciona información sobre el valor del programa educativo. La evaluación formativa permite identificar diferentes aspectos del aprendizaje a lo largo del proceso o de la intervención educativa, con el fin de profundizar o modificar el aprendizaje subsecuente, valorando los atributos tanto positivos como negativos; usualmente se realiza durante la implementación de un proyecto o programa, o durante el desarrollo de un curso; es fundamental en este tipo de evaluación que se brinde una realimentación a los estudiantes. De acuerdo con algunos autores, la *evaluación diagnóstica* representa un tipo de evaluación formativa, la cual permite determinar un punto de partida del aprendizaje que posee el estudiante; se aplica al inicio de un curso o de una unidad temática, y permite ajustar el programa de aprendizaje.

Tipos de evaluación de acuerdo con quién la realiza. Esta clasificación tiene dos componentes: la *evaluación interna* o *basada en la escuela* y la *evaluación externa* o *estandarizada.* La evaluación interna suele ser elaborada por el profesor de los estudiantes, usualmente en colaboración con los mismos estudiantes, y se desarrolla tanto durante el curso como al final de alguna unidad temática o ciclo escolar. La evaluación externa es diseñada fuera de las escuelas para asegurar que las preguntas, las condiciones para la aplicación, métodos de calificación e interpretaciones sean consistentes y comparables entre los estudiantes; generalmente intervienen grupos de pares, comités de expertos u organismos especializados en cuestiones de evaluación.

Tipos de evaluación de acuerdo con la interpretación de resultados. Dentro de esta clasificación se encuentran la *evaluación referida a la norma (relativa)* y la *evaluación referida al criterio (absoluta)*[10]. En la referida a la norma, la evaluación indica los resultados obtenidos por los alumnos y la relación con el estándar basado

en el desempeño de una gran muestra externa representativa de estudiantes (la norma o promedio); el estándar determinado por este tipo de evaluación puede variar dependiendo del desempeño del grupo promedio. En la evaluación referida al criterio, el proceso refleja la cantidad de conocimientos que el estudiante sabe en realidad, sin necesidad de compararlo con otros; este tipo de evaluación se asocia fuertemente con una educación basada en competencias o en contenidos, en este caso, el estándar se fundamenta en el conocimiento y las habilidades que el estudiante debe poseer para acreditar un curso.

1.3 Instrumentos de evaluación

Un instrumento de evaluación se define como una herramienta que se elige o se construye para medir o valorar aspectos o características identificados en el proceso de evaluación[5]. Existe una gran variedad de instrumentos que permiten medir el aprendizaje obtenido por los estudiantes. Un examen (o prueba) es un tipo de instrumento de evaluación utilizado para medir o cuantificar el logro de un determinado objetivo de aprendizaje[11].

Diaz-Barriga[12] menciona tres tipos diferentes de técnicas e instrumentos de evaluación, los cuales son:

Técnicas informales. Se utilizan dentro de episodios de enseñanza con una duración breve. No suelen ser presentados a los alumnos como métodos de evaluación, por lo que suelen ser muy útiles para valorar sus desempeños. Dentro de estas técnicas se pueden identificar la observación de las actividades realizadas por los alumnos y la exploración por medio de preguntas formuladas por el profesor durante la clase.

Técnicas semiformales. Requieren un mayor tiempo de preparación que las informales, demandan mayor tiempo para su valoración y exigen respuestas más duraderas por parte de los alumnos. Algunos tipos de esta variante son los trabajos y ejercicios que el estudiante realiza durante la clase, las tareas y trabajos que los

profesores encomiendan para su realización fuera del horario de clases y la evaluación de un portafolios.

Técnicas formales. Estas técnicas exigen un proceso de planeación y elaboración sofisticados y suelen aplicarse en situaciones que demandan un mayor grado de control. Son percibidas por profesores y estudiantes como una situación verdadera de evaluación. Algunas técnicas usadas en esta categoría son las pruebas o exámenes, los mapas conceptuales, o la evaluación del desempeño (con instrumentos como las rúbricas o las listas de cotejo)

Como hace notar George Miller[13], no existe un método de evaluación que pueda ofrecer la suficiente información para emitir un juicio completo sobre el desempeño profesional del médico. Para definir los instrumentos adecuados para la evaluación de los diferentes niveles de conocimiento, este autor propone un modelo teórico conocido actualmente como “Pirámide de Miller” para clasificar los escenarios en los que la evaluación puede ocurrir. En la base de la pirámide se encuentra el *conocimiento*, en este nivel es necesario asegurarse que el estudiante *sabe* lo necesario para llevar a cabo sus funciones profesionales de manera efectiva. El segundo nivel se refiere a que el estudiante debe *saber cómo* utilizar el conocimiento para lograr el objetivo, desarrollando la habilidad de adquirir información de diversas fuentes, analizarla e interpretarla, para finalmente trasladar los hallazgos en un diagnóstico o plan de manejo (donde el autor cita a Webster, para quien esta es la definición de *competencia*). En el tercer nivel, es importante que el alumno sea capaz de *demostrar cómo* es su *desempeño*, pues no es suficiente con solamente saberlo o saber cómo hacerlo. Finalmente, la punta de la pirámide se refiere al hecho de que el estudiante sea capaz de *hacer* lo que fue demostrado en los entornos artificiales en su ejercicio profesional independiente, es el componente de la *acción* en el comportamiento profesional.

Cada nivel de esta pirámide requiere de instrumentos de evaluación más complejos, ya que las actividades o procesos a evaluar son cada vez de más alto

nivel. Para el primer nivel de la pirámide, Wass *et al.*[1] proponen el uso de exámenes con reactivos de opción múltiple y otros instrumentos que ofrecen una gran confiabilidad para evaluar el conocimiento del estudiante, como las preguntas de verdadero y falso, emparejamiento, entre otras; en el segundo nivel de la pirámide, en donde se evalúa el conocimiento del procedimiento, los autores proponen el uso de preguntas tipo ensayo modificadas o problemas de manejo de pacientes, que permiten evaluar procesos cognitivos de más alto nivel. La evaluación del desempeño se puede realizar con diversos métodos, como el uso de casos clínicos cortos o largos con el apoyo de pacientes no estandarizados (que cuenta con una baja confiabilidad), el examen clínico objetivo estructurado, y el uso de pacientes estandarizados o simuladores, para demostrar la competencia clínica antes de enfrentarse a los pacientes reales. La evaluación del desempeño profesional (el *hacer* de la pirámide) ha representado un reto en la educación médica, pues requiere de la observación del estudiante en la situación real con un paciente; una estrategia planteada es el uso del portafolio para evaluar el desempeño del estudiante al final del año, más que una evaluación sumativa. La Pirámide de Miller y los diferentes instrumentos de evaluación para cada uno de los niveles de acuerdo con la propuesta de Wass *et al.* se muestran en la **Figura 1.1**.

La evaluación de los niveles más bajos de la pirámide requiere una definición operativa del concepto de “saber” (“*knows*”) y “saber cómo” (“*knows how*”). Ambos niveles se pueden englobar en el concepto de “conocimiento cognitivo”, que de acuerdo con Downing[14], se refiere al aprendizaje asociado con alguna teoría o habilidad mental. Los exámenes escritos son un instrumento adecuado para la evaluación de estos niveles de la pirámide, de los cuales podemos describir dos formatos diferentes: los que requieren la construcción de una respuesta y aquellos en los que se requiere seleccionar una respuesta correcta[15]. Dentro de los reactivos de selección, uno de los formatos más frecuentes son los reactivos de opción múltiple.

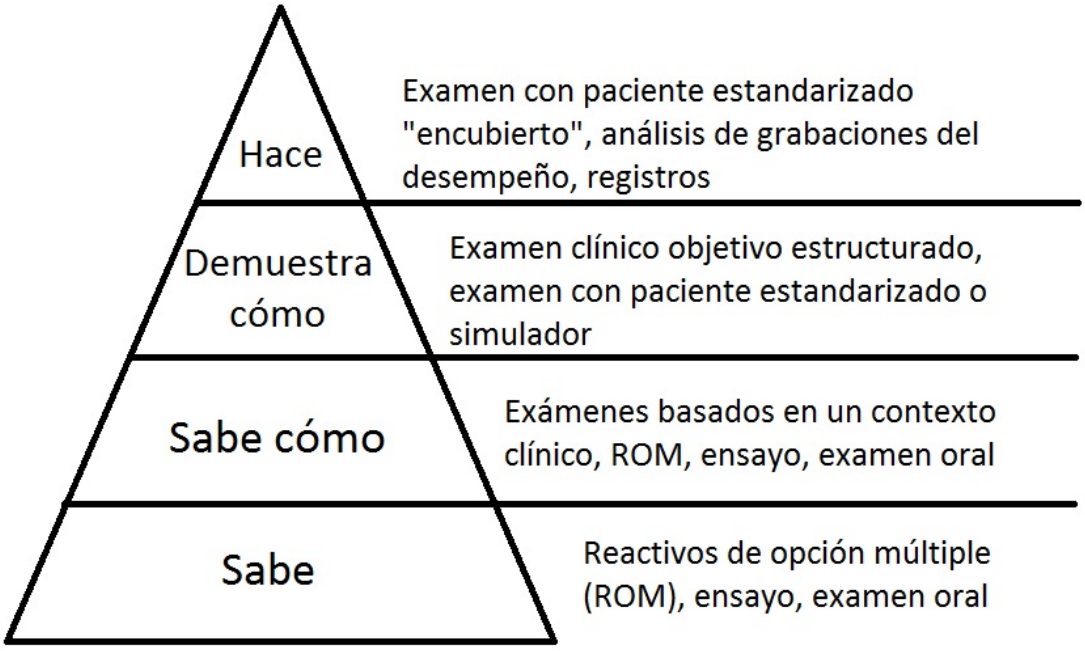


Figura 1.1 Pirámide de Miller (modificado de Wass V *et al, 2001*)

1.4 Reactivos de opción múltiple

Un reactivo de opción múltiple (ROM) se puede definir como “un problema o planteamiento que debe resolverse; presenta varias opciones de respuesta estructurada, de las cuales sólo una es correcta”[16]. Los ROM pueden utilizarse para evaluar distintos resultados de aprendizaje, usualmente conocimiento, comprensión y aplicación, de acuerdo con la taxonomía de objetivos educativos de Bloom[17].

Los reactivos de opción múltiple convencionales están constituidos por los siguientes elementos fundamentales[18]:

Tallo. También llamado base o enunciado, este elemento está constituido por una pregunta, afirmación, enunciado o gráfico, acompañado de una instrucción que plantea explícitamente un problema. El tallo de un reactivo de opción múltiple puede contener una pregunta, establecer un enunciado incompleto, establecer un hecho, describir una situación, o una combinación de los anteriores[19].

Opciones de respuesta. Son alternativas de respuesta a la base o tallo. Existen dos tipos de opciones de respuesta:

- Opción correcta: la cual da solución al problema o completa el enunciado. También se conoce como "clave".
- Opciones incorrectas. También se conocen como distractores. Estas opciones corresponden a alternativas al tallo que deben ser incuestionablemente respuestas incorrectas[18].

Existen otros tipos de reactivos, también considerados de opción múltiple, pero que varían en su estructura, estos son los reactivos de tipo verdadero-falso, respuesta alternativa, emparejamiento, opción múltiple compleja, verdadero-falso múltiple, entre otros[20].

Los ROM tienen una gran cantidad de ventajas, por lo que son preferidos por muchos expertos en evaluación [21,22,23,24]; algunas son:

- Los ROM generan interpretaciones de los resultados con mayor validez de contenido, al incluir una muestra representativa de un mayor número de contenidos.
- La confiabilidad de los resultados es muy alta con suficientes ROM de buena calidad.
- Los ROM pueden ser probados antes de su aplicación, almacenados, usados, reutilizados, con la ventaja de un sistema de almacenamiento computarizado de bajo costo.
- Es posible la obtención de resultados objetivos y de manera rápida.
- Las teorías de los test (teoría clásica de los test, teoría de respuesta al ítem, teoría de la generalizabilidad) se adecuan fácilmente a respuestas binarias.
- La mayor parte de los tipos de contenido se pueden evaluar con este formato, incluyendo algunos tipos de pensamiento de alto nivel.

Así como tienen ventajas, el uso de reactivos de opción múltiple también tiene una serie de limitaciones[22,23,24]:

- Construir reactivos de calidad requiere de un tiempo considerable.
- Suele ser difícil encontrar distractores plausibles.
- Son inefectivos para medir algunos tipos de solución de problemas y la capacidad de organizar y expresar ideas.
- Los problemas en el mundo real se solucionan de manera diferente (proponer una solución y no seleccionar una solución de una lista de alternativas).
- Los resultados pueden verse influidos por la habilidad de lectura del sustentante.
- Existe una falta de realimentación a los procesos cognitivos individuales, ya que es difícil definir por qué un estudiante elige la respuesta incorrecta.
- Frecuentemente se examina la información factual sin evaluar niveles superiores de pensamiento.
- Suele haber más de una respuesta correcta “defendible”.
- Puede favorecer la adivinación de la respuesta correcta.

1.5 Características cualitativas de los reactivos de opción múltiple

Existen varios conceptos importantes que se deben de considerar al hablar de las características cualitativas de un instrumento de evaluación.

1.5.1 Validez

La validez se refiere a la evidencia que es presentada para apoyar o refutar el significado o la interpretación atribuida a los resultados de la evaluación[25]. De acuerdo con el documento de *Standards for educational and psychological testing*, la validez se refiere al grado en el que la evidencia y la teoría apoyan las interpretaciones de los resultados de los exámenes implicadas en el uso propuesto del examen[4]. En palabras coloquiales, la validez se refiere a que el examen mida lo que se supone que debe de medir[26]. La evaluación por sí misma no es “válida” o

"inválida", la validez no es tampoco una propiedad intrínseca del instrumento, sino de las inferencias que se hacen sobre los resultados obtenidos en la evaluación (tienen mayor o menor validez), en el momento particular en el que la evidencia de validez fue recabada.

La validez es tal vez el aspecto más importante que determina la importancia del uso de un instrumento de evaluación y las inferencias que se hacen sobre los resultados obtenidos. Convencionalmente se han descrito tres tipos de validez (de contenido, de criterio y de constructo)[27] y la evidencia de validez en el proceso depende de la suma total de estas evidencias. Al darse cuenta de que todos estos tipos de validez están relacionados con la obtención de evidencias de validez relacionadas con el constructo, se optó por un marco que unifica todas estas en la validez de constructo[28]. De acuerdo con Messick[29], existen diferentes fuentes de evidencia de validez, y estas se pueden clasificar en cinco tipos diferentes:

Contenido. Se refiere a las características de la prueba de acuerdo con una tabla de especificaciones, que permite describir las categorías y clasificaciones del contenido y proporciona la cantidad de reactivos de cada categoría así como los niveles cognitivos a evaluar. Este punto es tal vez una de las principales fuentes de validez en una prueba escrita. Uno de los aspectos importantes que surge aquí como evidencia de validez es la calidad de las preguntas de la prueba, que será abordado más adelante en este apartado de características cualitativas de los reactivos.

Proceso de respuesta. Se refiere a la evidencia que surge al eliminar todas las fuentes de error asociadas al proceso de aplicación del examen. De igual modo, se incluyen aspectos relacionados con la calificación del examen, desde el registro de las respuestas, la determinación de un estándar de pase, entre otros.

Estructura interna. Esta fuente de evidencias de validez está relacionada con el análisis de ítems (que será comentado en el siguiente apartado de características

cuantitativas de los ROM). Estas evidencias incluyen la dificultad, la discriminación, el error estándar de medición, entre otros parámetros.

Relación con otras variables. La correlación que tienen los resultados del examen con la medición de otra variable "criterio" aporta evidencia de validez. Esta fuente de evidencia de validez corresponde a lo que anteriormente se conocía como "validez de constructo", y sigue incluyendo los aspectos de validez concurrente y validez predictiva con fines de aportar evidencia que apoye las inferencias de los resultados.

Consecuencias. Se refiere al impacto que las puntuaciones, las decisiones y los resultados de la prueba tienen sobre los examinados, así como su impacto en los procesos de enseñanza y aprendizaje. Si el fin de la evaluación ofrece mayores beneficios que perjuicios es una fuente importante de validez de constructo y, por lo tanto, de la validez de la prueba. La definición de constructo se enuncia a continuación.

1.5.2 Constructo

Un constructo es un nombre que se asigna a una propiedad, también conocida como "variable latente"[30]. Downing y Haladyna[26] lo definen como una colección intangible de conceptos y principios abstractos, que se infieren del comportamiento y son explicados por la teoría psicológica o educativa. Debido a estas características, se requiere la validación de los datos obtenidos en la evaluación, que permita relacionar la interpretación de los resultados de esta en una compleja red de teoría, hipótesis y lógica, que es presentada para apoyar o refutar qué tan razonables son las interpretaciones deseadas. Como lo define Messick[29], la validez de constructo se basa en la integración de la evidencia que apoya la interpretación de los resultados del examen.

1.5.3 Tabla de especificaciones

Dentro de las evidencias de validez relacionadas con el contenido se encuentra el hecho de que las preguntas que forman parte de una prueba estén fundamentadas en un documento, como es la tabla de especificaciones. Este documento (conocido en inglés como *test blueprint*) es una tabla que permite realizar una alineación entre los objetivos, la instrucción o didáctica y la evaluación[31]. Si bien existen diferentes propuestas para su elaboración, algunos de los principales componentes que deben incluirse son[32]:

1. Relación entre los objetivos o metas seleccionados para el examen
2. Relación entre los niveles de aprendizaje
3. Estructura del examen
4. Número total de reactivos
5. Número total de reactivos para cada objetivo y nivel de aprendizaje
6. Habilidades seleccionadas a partir del perfil de referencia

El uso de una tabla de especificaciones representa evidencia de validez relacionada con el contenido y con el proceso de respuesta[33]. En la primera, la relación de la tabla de especificaciones con el perfil de referencia utilizado para definir el contenido de la prueba enfatizan la congruencia que existe entre ambos documentos, mientras que para la segunda está apoyada por la definición del proceso de aplicación del instrumento a partir de las características definidas previamente para la estructuración del examen.

1.5.4 Recomendaciones para la elaboración de reactivos de opción múltiple

Como se mencionó previamente en las evidencias de validez relacionadas con el contenido de la prueba, es importante considerar dentro de las características cualitativas de los ROM su adecuada elaboración. En 1989, Haladyna y Downing[24] realizaron un análisis de 46 fuentes bibliográficas sobre medición educativa, y

propusieron una clasificación con 43 reglas para la elaboración de reactivos de opción múltiple; esta clasificación incluía las siguientes categorías:

1. Escritura del reactivo
    a. Procedimientales – 7 criterios
    b. Relacionados con el contenido – 10 criterios
2. Construcción del tallo – 6 criterios
3. Desarrollo de las opciones – 12 criterios
    a. Opción correcta – 2 criterios
    b. Distractores – 6 criterios

Posteriormente, los mismos autores junto con Rodríguez[20] realizaron una revisión de esta taxonomía, analizando 27 libros de texto y 27 estudios de investigación publicados desde 1990, en donde proponen una nueva guía para la elaboración de reactivos de opción múltiple (**Anexo 1**). Esta nueva clasificación incluye las siguientes categorías:

1. Aspectos del contenido – 8 criterios
2. Aspectos del formato – 2 criterios
3. Aspectos del estilo – 3 criterios
4. Escritura del tallo – 4 criterios
5. Escritura de las opciones – 14 criterios

Varios autores también han propuesto diversas metodologías para la adecuada elaboración de ROM. Vale la pena destacar el trabajo de Moreno *et al.*[34], quienes hacen una modificación de la metodología propuesta por Haladyna, Downing y Rodríguez, proponiendo un total de 12 directrices para la elaboración de reactivos de opción múltiple, la cual incluye las siguientes categorías:

1. Elección del contenido que se desea evaluar – 2
2. Expresión del contenido en el reactivo – 3
3. Construcción de las opciones – 7

Una de las recomendaciones que propone Haladyna involucra el hecho de "evaluar aprendizaje del nivel más alto". En el campo de la psicología, Benjamin Bloom propuso en 1956 una taxonomía para clasificar los objetivos educativos en el dominio cognitivo, los cuales permiten lograr una congruencia entre estos y las actividades de enseñanza y evaluación descritas en el currículo, tener un lenguaje común sobre las metas de aprendizaje que facilita la comunicación entre los participantes del proceso educativo, servir como base para determinar el significado específico de los objetivos educativos globales, así como brindar un panorama de las posibilidades educativas contra las que se puede contrastar la profundidad y amplitud de un curso o programa[35].

La taxonomía original de Bloom posee un total de 6 categorías, cada una de las cuales implica un dominio cognitivo en particular, jerarquizadas de lo simple a lo complejo y de lo concreto a lo abstracto, asumiendo también que era una clasificación acumulativa, en donde cada uno de los niveles implican la adquisición y dominio de los niveles inferiores a él.

En la literatura médica, se han hecho diversas adaptaciones a esta taxonomía[35,36,37,38], en donde en la mayoría se expresan tres niveles diferentes del dominio cognitivo del aprendizaje para su aplicación en la elaboración de reactivos de opción múltiple (**Tabla 1.1**).

Tabla 1.1. Diferentes definiciones de los niveles taxonómicos del dominio cognitivo de Bloom.

Fuente	Nivel 1	Nivel 2	Nivel 3
Krathwohl DR.[35]	Evocar conocimiento relevante de la memoria a largo plazo.	Determinar el significado de mensajes instruccionales, incluyendo comunicación oral, escrita y gráfica.	Llevar a cabo o utilizar un procedimiento en una situación determinada.
Palmer EJ, Devitt PG.[36]	Recuerdo de la información.	Entender y ser capaz de interpretar la información.	Uso del conocimiento y la comprensión en nuevas circunstancias.
Buckwalter JA, Schumacher R, Albright JP, Cooper RR.[37]	Reconocimiento y recuerdo de información aislada.	Comprensión e interpretación de la información. Demostrar la comprensión de la información haciendo uso de ella. El uso de la información puede ser una extrapolación de la información o la explicación de esta en otra forma (verbal, tabular, visual, morfológica o gráfica).	Aplicación del conocimiento en la solución de un problema específico. Demostrar el análisis del problema, recordar información y principios relevantes, y aplicar la información y los principios en la solución de un problema específico.
McGuire C.[38]	Recuerdo de información aislada: Reconocimiento de lesiones morfológicas y preguntas sobre hechos, conceptos, principios, procesos y teorías específicas. Usualmente responde a preguntas como "¿Qué es X?"	Selección de una generalización relevante para explicar fenómenos específicos. Difieren del nivel I porque preguntan "¿Por qué es X verdadero?" en lugar de "¿X es verdadero?"	Interpretación de la información: Explicar información verbal, tabular, morfológica o gráfica de otra manera, o hacer intrapolaciones o extrapolaciones de la información. Aplicar un principio o combinación de principios en una situación. El contenido específico del problema es nuevo para el estudiante, pero el problema involucra un patrón familiar.

1.6 Características cuantitativas de los reactivos de opción múltiple

1.6.1 Teorías de los tests

Las teorías estadísticas de los tests permiten asegurar que las inferencias hechas a través de las pruebas sean adecuadas y pertinentes, estimando así la validez y confiabilidad de los mismos[39]. Para lograr que el error en estas mediciones se reduzca al mínimo y valorar con precisión las puntuaciones o resultados obtenidos, es necesario utilizar modelos estadísticos. Existen dos enfoques principales para el análisis psicométrico: la Teoría Clásica de los Tests (TCT) y la Teoría de Respuesta al Ítem (TRI).

Teoría clásica de los tests (TCT). Modelo propuesto por Spearman a principios del siglo XX, el cual consiste en asumir que la calificación obtenida por un sujeto en un test (puntuación empírica o X) está conformado por dos componentes, la puntuación verdadera de esa persona en ese test (V), la cual es una constante para cada persona, y el error de medida que, como con cualquier instrumento de medición, se comete al medir el rasgo con ese test (e)[40]. Esto se puede expresar de la siguiente manera:

$$X = V + e$$

Varios supuestos se tienen que cumplir bajo el modelo clásico de esta teoría[41]:

- La puntuación verdadera y el error de medición no tienen correlación.
- El valor del error promedio de la población de evaluados es cero.
- El valor del error en pruebas paralelas no tiene correlación.

Para el análisis de ítems con el uso de esta teoría se puede utilizar software especializado, como pudiera ser CITAS, jMetrik, Iteman, entre otros.

Teoría de respuesta al ítem (TRI). El modelo TRI determina el comportamiento de una persona respondiendo a un reactivo en particular. Se basa en el supuesto de que existe una relación funcional entre los valores de la variable que miden los reactivos y la probabilidad de acertarlos (en una función denominada Curva

Característica del Ítem)[39,41]. Si bien la TRI da respuesta a varias limitaciones que muestra la teoría clásica, esta última suele preferirse debido a la sencillez relativa de sus procedimientos y los supuestos en los que se fundamenta, así como la utilidad práctica que se le ha demostrado con el paso de los años. Para el análisis psicométrico bajo este modelo se puede utilizar software especializado, como Facets, jMetrik, BMIRT, entre otros.

1.6.2 Análisis de ítems

El análisis de ítems se refiere a los procedimientos dirigidos a extraer información sobre la calidad de los reactivos[40]. En general, se refiere únicamente al análisis cuantitativo de los mismos. Algunos de los parámetros que se reportan en este análisis son:

Confiabilidad. También llamada fiabilidad, se refiere al grado en que un test o prueba puede ser replicado o reproducido[9]. Representa un elemento esencial de la validez. Existen diferentes métodos para estimar la confiabilidad[42]:

- Test-retest. Consiste en que el individuo tome varias veces el mismo test, ya que de acuerdo con la teoría clásica, su verdadero score o resultado deberá ser el promedio de los resultados obtenidos en un gran número de tests.
- Pruebas paralelas. Consiste en la aplicación de dos versiones del mismo test, con reactivos distintos, pero que pretenden medir lo mismo.
- Consistencia interna. Expresan el punto en el que las respuestas son lo suficientemente coherentes (entre sí) para concluir que todos los reactivos miden lo mismo y que, por ello, son sumables para representar un mismo rasgo. Una de las fórmulas más utilizadas es el coeficiente α de Cronbach.

Índice de dificultad. Se refiere a la cantidad de examinados que contestan de manera correcta el reactivo: a mayor número de estudiantes con la respuesta correcta, el reactivo tiene menor dificultad, y viceversa. Dentro del marco de la TCT, el índice de dificultad depende de la población que es examinada. Se entiende

también como la facilidad de un reactivo[43]. Al ser este índice una proporción, se expresa en valores de 0 a 1. Un reactivo que es contestado por el 90% de los sustentantes tendrá un índice de dificultad de 0.9 (considerado un índice alto, es decir, un reactivo sencillo de contestar), mientras que un reactivo que es contestado por únicamente el 20% de los sustentantes tendrá un índice de dificultad de 0.2 (considerado un índice bajo, es decir, un reactivo muy difícil).

Discriminación. Se refiere a la tendencia de un reactivo a ser respondido de manera correcta por los alumnos que poseen el conocimiento o habilidad que se evalúa en él, y a ser respondido de manera incorrecta por aquél que no posee dicha habilidad o conocimiento. Existen varias maneras de calcular la discriminación del reactivo[40].

- Índice de discriminación. Se refiere a la diferencia entre la proporción de estudiantes del grupo de alto desempeño y del grupo de bajo desempeño que seleccionaron la respuesta correcta en el reactivo.
- Correlación biserial (Rbis). Se refiere a la correlación entre el valor obtenido en el reactivo (0 si es incorrecta la respuesta y 1 si es correcta) y la calificación total obtenida en la prueba.
- Correlación punto biserial (Rpbis). Se refiere también a la correlación entre el valor obtenido en el reactivo y la calificación total en la prueba. A diferencia de la correlación biserial, la Rpbis está diseñada para la correlación de variables dicotómicas, mientras que en la Rbis el análisis se hace para una variable continua que se dicotomiza para el análisis. Esta correlación puede tomar valores de 1 a -1, entre más se aleja este indicador del cero, quiere decir que existe una tendencia de que una respuesta sea elegida por el grupo de mayor o menor desempeño en la prueba. Un índice de discriminación cercano a cero implica que esa opción de respuesta aporta muy poca o nula información para diferenciar a los estudiantes de mayor o menor desempeño en el examen únicamente con esa información.

Dentro del análisis cuantitativo se puede incluir el análisis del comportamiento de los distractores. Un distractor no funcional (NFD por sus siglas en inglés, *non*

functioning distractors) se puede definir como aquella opción incorrecta que es elegida por menos del 5% de los sustentantes[44] (es decir, un índice P menor a 0.05. También existen distractores que tienen un comportamiento diferente al esperado (un Rpbis positivo, lo que quiere decir que el distractor tiende a ser elegido en mayor proporción por estudiantes de buen desempeño en la prueba con respecto a los estudiantes de mal desempeño); estos distractores arrojan error en el proceso de medición y es necesario considerar su pertinencia dentro del reactivo.

2. Antecedentes

2.1 La Facultad de Medicina y el Plan de Estudios 2010

"El Plan de Estudios 2010 de la Licenciatura de Médico Cirujano de la Facultad de Medicina de la Universidad Nacional Autónoma de México se inserta en un contexto caracterizado por vertiginosos cambios económicos, sociales, culturales, científicos y tecnológicos que influyen en la formación del médico del futuro"[45]. Desde el Plan Único de Estudios, que se puso en marcha en el ciclo escolar 1994-1995, se incorporan en su estructura el marco de referencia, la misión de la Facultad, el perfil profesional del egresado y la metodología educativa, así como los ejes del plan de estudios, la vinculación de las ciencias básicas con las clínicas y la formación integral de los estudiantes, tomando en cuenta los lineamientos o pautas para seleccionar y organizar las asignaturas, el perfil epidemiológico de la población, los conocimientos, habilidades y destrezas indispensables para la práctica de la medicina general, y la secuencia, niveles de complejidad e interacción de contenidos y asignaturas.

Después de establecer la Comisión de Evaluación del Plan Único de Estudios en el año 2004, se realizó la propuesta de modificación de este plan, que de acuerdo al referente contextual (que toma en cuenta las demandas, necesidades y retos que surgen del diagnóstico de salud, la organización de los servicios de salud, el estado actual y las tendencias futuras en la medicina y en la formación profesional), institucional (la organización académico-administrativa de la Facultad, así como las actividades docentes y de investigación) y curricular (los fundamentos curriculares, los currículos afines, los resultados de la evaluación del Plan Único de Estudios y la organización del currículo por asignaturas, con un currículo nuclear, un enfoque por competencias y la integración), permitió hacer la propuesta del Plan de Estudios 2010, el cual inició sus actividades en el ciclo escolar 2010-2011.

2.2 La asignatura de Inmunología en la licenciatura de Médico Cirujano de la Facultad de Medicina de la UNAM

En el año 1985 el H. Consejo Técnico de la Facultad de Medicina decidió que la materia de Bioquímica fuera nombrada “Bioquímica e Inmunología” y los contenidos de esta última fueron incorporados a esa asignatura[46]. Con la implementación del Plan Único de Estudios[47], y debido a los grandes avances en esta ciencia y la importancia que tiene en la formación del Médico Cirujano, desde el año 1996 se imparte como una asignatura independiente, en el segundo año de la carrera, con un total de 10 créditos y 120 horas académicas, a cargo del Departamento de Bioquímica de la Facultad. Desde esa fecha, el programa académico ha sufrido una gran variedad de cambios, debido a la gran velocidad con la que se genera nuevo conocimiento en esta área.

A partir de la modificación del plan de estudios, con la entrada del Plan 2010, la asignatura pasó a formar parte del área de conocimiento “Bases Biomédicas de la Medicina”, y cambió de impartirse de forma anual a semestral, lo que conllevó a una disminución en el número total de horas de clase (85 horas académicas) y de créditos (un total de 7)[45]. Esta reducción del tiempo para impartir la asignatura representó un reto para la Coordinación de Enseñanza de Inmunología y para sus profesores, pues el programa académico conserva los mismos temas y es sujeto de continuar sufriendo modificaciones conforme el conocimiento sigue aumentando, obligando a mejorar los procesos de enseñanza-aprendizaje-evaluación, debido a que es una asignatura indispensable en su formación como Médicos Cirujanos. Durante el año escolar 2012 (primer año en el que la asignatura fue impartida para el Plan 2010, debido a que está ubicada en el segundo año de la carrera), aproximadamente la mitad de la generación correspondió a estudiantes que habían recursado el año anterior una o más asignatura del primer año de la carrera, o que se encontraban recursando la misma asignatura de Inmunología, mientras que la otra mitad correspondía a estudiantes regulares inscritos en el nuevo plan de estudios. En el año escolar 2014 se decidió hacer nuevamente una modificación en el mapa curricular de la asignatura[48], volviendo a la modalidad anual para la

impartición de la asignatura con un total de 42 grupos en este nuevo plan, que incluye alumnos regulares y recursadores. La asignatura de Inmunología recibe anualmente más de mil estudiantes, los cuales son objeto de todos los procesos didácticos y de evaluación correspondientes a la impartición de la asignatura.

El temario de la asignatura de Inmunología consiste en una primera parte (Inmunología básica), en donde se revisan las características y mecanismos efectores del sistema inmune, y se conforma por cuatro unidades temáticas; en la segunda parte, se retoma todo este conocimiento se enfoca en un entorno clínico (Inmunología clínica), lo cual conforma la tercera unidad temática[49]. Debido a esto, el alumno tiene que aprender una nueva nomenclatura y una gran cantidad de conceptos en aquella primera parte, para que después pueda aplicarlos en situaciones a las que se enfrentarán en el ejercicio clínico. La estructura del programa académico consta, por lo tanto, de cinco unidades temáticas, que son:

1. Generalidades
2. Respuesta inmune innata
3. Respuesta inmune adaptativa
4. Regulación de la respuesta inmune
5. Introducción a la inmunología clínica e inmunopatología

Un aspecto importante del proceso educativo consiste en mejorar, de manera continua y permanente, los procesos de enseñanza-aprendizaje, pero para poder observar la calidad del aprendizaje logrado por el alumno tenemos que enfocarnos en el proceso de evaluación, el cual nos mostrará si se estos procesos se están llevando de manera adecuada.

2.3 Evaluación de la asignatura de Inmunología

Para poder asegurar que el alumno alcanza los objetivos propuestos en el programa académico y desarrolla las competencias necesarias para alcanzar el Perfil Intermedio I y a mediano plazo el perfil de egreso, es necesario realizar una

buena evaluación del aprendizaje, satisfaciendo los criterios de validez y confiabilidad necesarios para ello.

De acuerdo con los "Lineamientos generales para la evaluación de los alumnos en las asignaturas de la carrera de Médico Cirujano"[47], aprobado por el H. Consejo Técnico de la Facultad de Medicina, en el apartado 3 se establece que "en todas las asignaturas se contará con dos calificaciones: la del profesor y la departamental", en este mismo apartado, en el primer inciso se establece que la ponderación de cada una de ellas podrá ser del 40 al 60 por ciento, debiendo sumar un total de 100%; en el apartado 5 se menciona que "la evaluación departamental corresponderá a la calificación obtenida por el alumno en los exámenes teóricos y prácticos parciales"; se menciona también que "los exámenes serán elaborados colegiadamente y aplicados por los profesores del curso, bajo la coordinación de los departamentos o secretaría correspondientes). Estos lineamientos se continuaron aplicando durante los ciclos escolares 2012, 2013 y 2014 en las asignaturas de los dos primeros años de la licenciatura.

En el caso de la asignatura de Inmunología, el examen departamental tiene una ponderación del 50% sobre el total de la calificación del estudiante. Al ser estas evaluaciones departamentales un medio sumativo para definir un porcentaje importante de la calificación del estudiante, el instrumento utilizado en la asignatura son los exámenes con reactivos de opción múltiple (tanto los convencionales, de verdadero y falso y de correlación), de los cuales se incluyen entre 60 y 70 en cada uno de los exámenes parciales, mientras que en los ordinarios y extraordinarios se aplicaban un total de 102 reactivos, aunque desde el año escolar 2014 disminuyeron a 80 para este tipo de exámenes. Las características de cada reactivo tienen que ser congruentes con los objetivos y competencias que se espera que el alumno alcance o desarrolle, por lo que la evaluación de los reactivos es un punto esencial para saber si esto está sucediendo.

Durante la impartición de la asignatura en el Plan Único de Estudios, se realizaban 3 exámenes parciales departamentales. Durante el año escolar 2012 (semestre 2012-1), para el Plan de Estudios 2010 se realizaron también tres exámenes parciales; esto cambió para el año escolar 2013 (semestre 2013-1), ya que debido al poco tiempo para la impartición de la asignatura y la cercanía entre cada examen parcial, se decidió aplicar únicamente dos exámenes parciales departamentales, lo que se acompañó de un aumento en la cantidad de temas de cada uno de ellos. En el año escolar 2014 (semestres 2014-1 y 2014-2), se aprobó por el H. Consejo Técnico de la Facultad de Medicina un cambio curricular en la impartición de la asignatura, pasando nuevamente a una modalidad anual, por lo que se decidió regresar a la aplicación de 3 exámenes departamentales, uno para cada una de las unidades temáticas de la asignatura[50].

A partir de la creación de la Coordinación de Evaluación como parte de la Estructura Administrativa de los Departamentos Académicos de la Facultad de Medicina[51], se inició la estructuración de los documentos necesarios para formalizar el proceso de evaluación dentro de la asignatura de Inmunología, con lo que a partir de un trabajo colegiado se generó una primera propuesta de tabla de especificaciones, a partir de los objetivos temáticos que se encontraban plasmados en el programa académico de la asignatura (**Anexo 2**).

3. Método

3.1 Planteamiento del problema

Con la implementación del Plan de Estudios 2010, el total de horas curriculares para la impartición de la asignatura pasó de ser 120 a 85, lo que representó una disminución total de 35 horas, sin haber una modificación de los contenidos.

Durante el año escolar 2012 se realizaron 6 exámenes parciales sumativos de Inmunología, pues se encontraban inscritos tres grupos de alumnos diferentes: aquellos que habían recursado primer año y cursaban por primera vez la asignatura de Inmunología, aquellos alumnos que estaban recursando la asignatura de Inmunología, todos estos inscritos en el Plan Único de Estudios, y los alumnos que cursaban también por primera vez Inmunología, pero inscritos en el Plan de Estudios 2010. Esto sumaba un total de 1,031 alumnos inscritos en la asignatura durante este ciclo escolar (aproximadamente la mitad de cada Plan de Estudios). A partir del año escolar 2013, la totalidad de alumnos inscritos en la asignatura (1226 estudiantes) pertenecía al Plan de Estudios 2010, aunque con cambios en tanto en la forma de evaluación de la asignatura (sólo dos exámenes parciales para el año escolar 2013) y de nueva cuenta, la impartición de la asignatura en modalidad anual (a partir del año escolar 2014), con un total de 1139 estudiantes inscritos.

El Plan de Estudios 2010 contempla el Plan de Evaluación y Actualización del Plan de Estudios 2010[45], como una estrategia cuyo propósito es “mejorar la calidad de la educación médica y el funcionamiento del plan de estudios y los programas académicos”. Dentro de este Plan, se describe que la evaluación que se realiza en la Facultad de Medicina tiene que cumplir tres características: ser válida, confiable y objetiva.

La implementación en la Facultad de las Coordinaciones de Evaluación dentro de los Departamentos Académicos conlleva una serie de necesidades para llevar a

cabo de manera objetiva y sistemática los procesos de evaluación de cada asignatura, por lo que la elaboración de instrumentos que arrojen resultados con evidencia de validez suficiente es de vital importancia para una adecuada evaluación del aprendizaje de los estudiantes, y que este proceso permita obtener información adecuada para tomar decisiones que permitan mejorar el proceso educativo.

Las preguntas de investigación que surgen son: ¿Cuáles son las características cualitativas y cuantitativas de los reactivos utilizados en las evaluaciones parciales sumativas de la asignatura de Inmunología aplicadas en la Facultad de Medicina de la UNAM durante los años académicos 2012, 2013 y 2014? y ¿cuál es el impacto que estas diferencias tienen en la calificación obtenida por el estudiante?

3.2 Objetivos

Objetivo general

Evaluar las características cualitativas y cuantitativas de los reactivos de opción múltiple utilizados en la asignatura de Inmunología, en los exámenes realizados durante los años escolares 2012, 2013 y 2014 en la Facultad de Medicina de la UNAM.

Objetivos específicos

- Generar un instrumento con evidencia de validez para la evaluación de reactivos de opción múltiple.
- Analizar el comportamiento psicométrico de los reactivos, a partir del modelo clásico de la teoría clásica de los test.
- Identificar los reactivos con las características adecuadas para integrar el banco de reactivos de la Coordinación de Evaluación de Inmunología, así como aquellos que deben ser revisados y modificados.

- Analizar las diferencias en la calificación de los estudiantes, asociadas al uso de reactivos con errores en su elaboración.

3.3 Justificación

La evaluación de los reactivos permitirá obtener evidencias de validez de los exámenes departamentales que se aplican como parte de la evaluación sumativa de los estudiantes.

De acuerdo con los estándares[4], los exámenes deben diseñarse de modo que apoyen la validez de las interpretaciones realizadas acerca de los resultados de los mismos para los fines propuestos en la evaluación.

El diseño del instrumento contempla diferentes fuentes de evidencia de validez, de acuerdo a la propuesta de Downing[25], principalmente en relación con el contenido, ya que se contempla incluir la relación con la tabla de especificaciones como un criterio de validez, así como la calidad de los reactivos utilizados (de acuerdo a las recomendaciones propuestas en la literatura).

El análisis de ítems (características cuantitativas) forman parte de las evidencias de validez relacionadas con la estructura interna (modelo psicométrico utilizado y valores de dificultad y discriminación para los reactivos)[25], por lo que es necesario incluir estos resultados para la adecuada interpretación de los resultados de la prueba.

En la literatura se ha reportado que es muy frecuente encontrar reactivos que no cumplen con las recomendaciones para la elaboración de un buen reactivo. En un estudio realizado en una escuela de enfermería, Tarrant *et al.*[52] recolectaron 2770 preguntas de opción múltiple de diferentes exámenes aplicados durante cinco años, buscando 19 errores frecuentes reportados en la literatura, encontrando que más del 45% de los reactivos presentaban alguna violación a estas recomendaciones.

Si bien en la literatura se encuentran reportes sobre la identificación de errores en los reactivos de opción múltiple utilizados en exámenes sumativos, otros autores han demostrado las consecuencias de esto.

Pate y Caldwell[53] analizaron 187 reactivos del módulo de cardiovascular en una escuela de Farmacia en Luoisiana, utilizando como modelo las recomendaciones de Haladyna propuestas en 2002, identificando un total de 97 reactivos que no cumplían con al menos una de las recomendaciones (142 violaciones a estas recomendaciones en el total de reactivos); si bien esto coincide con el estudio de Tarrant, estos autores identificaron también que la dificultad de los reactivos que cumplía con los lineamientos tenía un promedio de respuestas correctas del 83.7%, en comparación con un 76.3% de los reactivos que no cumplían con alguna recomendación ($p=0.01$); la discriminación en este estudio (calculada a través de la correlación punto biserial) fue mayor para los reactivos que no cumplían con alguno de los lineamientos (0.255) con aquellos apegados a las recomendaciones (0.242).

Downing[54] realizó un estudio con reactivos utilizados en cuatro exámenes diferentes de asignaturas de los primeros dos años de la carrera de Medicina en la Universidad de Illinois, analizando un total de 219 reactivos y los patrones de respuesta de 637 estudiantes en esos mismos reactivos; las preguntas que quería responder fueron: 1) ¿Cuál es la incidencia de violaciones a las reglas de elaboración de reactivos (defectos) en estos exámenes, 2) ¿Qué efecto tienen estos defectos en la dificultad, discriminación y confiabilidad de la prueba?, y 3) ¿Qué efecto tienen estas violaciones en las decisiones de acreditación o reprobación de los estudiantes? El primer análisis arrojó resultados similares a los estudios previamente mencionados, con un total de 100 defectos encontrados en los diferentes reactivos. El segundo análisis también mostró, en general, una mayor dificultad en aquellos reactivos con defectos, con resultados variables en la discriminación y en la confiabilidad. El tercer análisis fue muy interesante, ya que al comparar el desempeño de los estudiantes que acreditaron el examen con reactivos

"estándar" con su desempeño en los reactivos defectuosos, se encontró que 102 de ellos reprobarían estos reactivos mal elaborados (también se demuestra que 30 estudiantes que reprobarían los reactivos normales hubieran acreditado con los reactivos defectuosos).

Estos estudios nos demuestran la importancia del análisis de los reactivos, tanto en sus características cuantitativas como cualitativas, desde el punto de vista de la necesidad institucional, como desde los aspectos éticos y de justicia en una evaluación sumativa, que en el caso de la Facultad de Medicina tiene un gran peso en la calificación de los estudiantes y que, en gran medida, marca su trayectoria escolar.

A partir de la implementación de los cambios curriculares que ha tenido la asignatura de Inmunología con el Plan de Estudios 2010, se ha visto un aumento en el índice de reprobación de los estudiantes, por lo que es importante contar con evidencias de validez suficientes sobre los instrumentos que se utilizan para la evaluación del aprendizaje del estudiante; en este trabajo el estudio de estas evidencias se centra en los reactivos de opción múltiple utilizados como parte de la evaluación departamental de la asignatura.

3.4 Diseño de investigación

El diseño del estudio es no experimental, de tipo descriptivo y correlacional, transversal y retrospectivo. Se llevó a cabo con reactivos utilizados en los exámenes parciales sumativos de la asignatura de Inmunología, en la Facultad de Medicina de la Universidad Nacional Autónoma de México.

3.5 Selección de la muestra

Los reactivos que serán objeto de análisis comprenden aquellos utilizados en los años escolares 2012, 2013 y 2014. Debido a la uniformidad de la población que

presenta los exámenes parciales (que representa a la totalidad de la generación que cursa la asignatura) se decidió realizar la evaluación de los reactivos que se utilizaron únicamente en los exámenes parciales, quedando excluidos los que fueron aplicados en los exámenes ordinarios y extraordinarios. El total de reactivos seleccionados evalúan todos los temas que conforman el programa académico de la asignatura de Inmunología.

Como ya se mencionó previamente, durante el año escolar 2012 se impartía la asignatura en las modalidades de ambos planes de estudio (Único y 2010), lo que conllevó a la aplicación de 6 exámenes parciales de la asignatura (tres para cada plan de estudios), de los cuales los del Plan 2010 se aplicaron durante el primer semestre del año escolar, mientras que los del Plan Único se aplicaron distribuidos durante todo el año; cada uno de estos exámenes tuvo un total de 70 reactivos. Durante el año 2013, únicamente se impartió la asignatura en el Plan 2010, en su modalidad semestral, y se modificó la evaluación para aplicar únicamente dos exámenes parciales, cada uno con 70 preguntas. Finalmente, durante el año 2014 se impartió la asignatura en el Plan 2010 nuevamente en la modalidad anual, lo que permitió nuevamente la aplicación de 3 exámenes parciales, en este caso el primero con 60 reactivos y el segundo y tercero con 70 cada uno. En la **Tabla 3.1** se esquematizan estas cifras para facilitar su comprensión.

Los criterios de inclusión para los reactivos que serán objeto de evaluación en este estudio son únicamente reactivos de opción múltiple de cuestionamiento directo, que estén formados por un tallo y cuatro o cinco opciones de respuesta. Los criterios de exclusión son aquellos reactivos con formato de emparejamiento o de verdadero-falso.

Del total de reactivos contemplados, únicamente 308 cumplían con los criterios de inclusión, ya que el resto eran reactivos en formato de emparejamiento y de verdadero y falso.

Tabla 3.1 Distribución de reactivos, duración de la asignatura y número de exámenes aplicados en los periodos a analizar

Plan de Estudios	Duración de la asignatura	Periodo de impartición (semestres)	Número de exámenes parciales	Reactivos por examen / Reactivos por año
Único	Anual	2012-1 y 2	3	**70 / 210**
2010	Semestral	2012-1	3	**70 / 210**
2010	Semestral	2013-1	2	**70 / 140**
2010	Anual	2014-1 y 2	3	**60-70 / 200**
		Total:	**11**	**760**

3.6 Recolección de datos

Los exámenes a utilizar se obtuvieron a través de la Coordinación de Evaluación de Inmunología en el Departamento de Bioquímica de la Facultad, con autorización del Jefe de Departamento. La mayoría se encuentran en formato físico (el mismo examen que fue aplicado a los estudiantes), mientras que los exámenes que fueron aplicados en el sistema de exámenes en línea de la Facultad de Medicina (FM) se encuentran en el sistema de captura de los mismos, a los que se podrá acceder para su evaluación cuando lo sea requerido.

Se solicitó a la Secretaría de Servicios Escolares de la FM el envío de los archivos correspondientes a cada uno de los exámenes previamente mencionados que incluyen:

a) Las calificaciones de los estudiantes en documentos con extensión pdf.
b) Las bases de datos en documentos con extensión .xls, que contienen las respuestas de los estudiantes en cada uno de los exámenes (por opción de respuesta y por aciertos y errores).

c) Los análisis psicométricos generados con el programa Iteman.

Las bases de datos de los exámenes aplicados con el sistema de exámenes en línea de la Facultad también fueron proporcionadas por la Coordinación de Evaluación de Inmunología.

Estos archivos representan una primera evaluación cuantitativa de los reactivos (a partir del análisis psicométrico), pero es necesario volver a realizar estos análisis, ya que en casi todos ellos se han incluido resultados de estudiantes que no presentaron el examen (cuyo registro de respuestas se encuentra en blanco) y estos resultados modifican el comportamiento psicométrico de los reactivos.

3.7 Evaluación cualitativa de los reactivos

El análisis cualitativo comprende principalmente la adecuada elaboración de un reactivo, de acuerdo con las recomendaciones que se mencionan en la literatura. Para poder realizarlo, se realizó un instrumento que permitiera evaluar estas características en los reactivos analizados en este estudio. El proceso de elaboración del instrumento se describe a partir de las fuentes de evidencia de validez obtenidas para su uso propuesto y se describe en tres etapas.

3.7.1 Evidencia de validez relacionada con el contenido

Se realizó una revisión de diferentes autores que proponen lineamientos para la adecuada elaboración de un reactivo. Se tomó en primer lugar el modelo de Haladyna, Downing y Rodríguez[20], a partir de las traducciones realizadas en dos fuentes distintas. Con estos lineamientos se realizó un cuadro comparativo, tomando cinco fuentes diferentes de artículos[17,34,52] y libros especializados en la elaboración de reactivos para las ciencias de la salud[(16)(22)(55)], a partir de la cual se realizó, en primer lugar, el ordenamiento de las diferentes recomendaciones, al contemplar diferentes traducciones de la taxonomía de Haladyna[34,55], y agrupar

aquellos lineamientos que hicieran referencia al mismo aspecto (**Anexo 3**). Posteriormente se seleccionaron aquellos reactivos que resultaron más importantes y que tuvieron sentido en el proceso de evaluación sumativa que se lleva a cabo en la Facultad. Estos pasos permitieron generar una propuesta de instrumento en el formato de una lista de cotejo (**Anexo 4**), compuesto por 24 ítems que permiten evaluar si el reactivo tiene o no las diferentes características que son deseables en un buen reactivo de opción múltiple. En este instrumento se solicitó también que el evaluador identificara el nivel taxonómico (de acuerdo con la taxonomía de Bloom) del reactivo que evalúa.

Para el proceso de obtención de evidencias de validez del instrumento se decidió utilizar la validación de contenido por jueces. Los cinco participantes eran parte de la Secretaría de Educación Médica de la Facultad de Medicina, quienes cuentan con estudios de posgrado en el área educativa, así como con una amplia experiencia en la elaboración de exámenes y de reactivos de opción múltiple. El proceso de validación se realizó a través de la aplicación del instrumento en diez reactivos de opción múltiple que se han elaborado para su aplicación en el examen profesional de la licenciatura de Médico Cirujano de la UNAM; estos reactivos ya han sido previamente identificados como reactivos “buenos” o “malos”, de acuerdo con las características en su elaboración (basados empíricamente en diferentes recomendaciones); los reactivos que se utilizaron para la validación pertenecían a las asignaturas de Psicología Médica, Farmacología, Salud Pública y Embriología, dado que tienen mayor similitud con la aplicación inicial que se le dará al instrumento, al ser estas asignaturas de los primeros dos años de la carrera, al igual que la asignatura de Inmunología. Este proceso se llevó a cabo en un entorno virtual de aprendizaje (la plataforma Moodle versión 2.8), a través de la actividad “cuestionario”, que permitió que el grupo de expertos respondieran de manera más sencilla el instrumento, al identificar aquellos ítems que resulten confusos, redundantes, obsoletos, y principalmente, que puedan generar propuestas para mejorar la calidad del instrumento (**Anexo 5**). Estas propuestas fueron revisadas nuevamente en conjunto con otras fuentes[16,17,22,56] para definir la versión final del

instrumento para continuar con el proceso de obtención de evidencias de validez. Este instrumento evalúa 22 criterios, el primero relacionado con el nivel taxonómico del reactivo y el resto con los criterios de calidad propuestos en la literatura consultada (**Anexo 6**)

3.7.2 Evidencia de validez relacionada con el proceso de respuesta

El instrumento que resultó del proceso de validación por expertos fue aplicado por los profesores para evaluar los reactivos que fueron objeto del estudio, con el fin de evaluar las características cualitativas de los reactivos, así como también para aportar evidencia de validez del uso del instrumento, en este caso relacionada con el proceso de respuesta.

En esta etapa participaron un total de nueve profesores de la asignatura. De éstos, cinco cuentan con estudios de posgrado en la disciplina y cuatro con estudios de licenciatura en Medicina; cinco de ellos son profesores de carrera con más de 10 años de experiencia en la impartición de la asignatura y estudios de posgrado en el área (maestría o doctorado) y 4 de ellos son profesores de asignatura con una experiencia de entre 3 y 8 años en su impartición y estudios de licenciatura.

Para capacitar a los profesores de Inmunología en el uso del instrumento, se impartió durante los meses de marzo y abril de 2013 el “Curso-Taller Elaboración de reactivos para la evaluación del aprendizaje de Inmunología”, con una duración de 20 horas, en una modalidad *blended learning.* Dentro del curso se abordaron temas teóricos sobre evaluación, evaluación del aprendizaje, instrumentos de evaluación, análisis de ítems, pero se hizo especial énfasis en el perfil de referencia de la asignatura y las tablas de especificaciones para la elaboración de exámenes, así como también en las características de un buen reactivo de opción múltiple (tomando como modelo las recomendaciones de Haladyna), presentando el instrumento para la evaluación de reactivos y explicando cada uno de los ítems y realizando ejercicios prácticos para familiarizarlos con el instrumento (**Anexo 7**).

Dentro de las fuentes de evidencia de validez propuestas por Downing se menciona la adecuada capacitación de los profesores en la elaboración de reactivos[25], por lo que la elaboración de reactivos por parte de los profesores que tomaron el curso aporta validez a la calidad de las preguntas que conformen los exámenes.

Una vez capacitados los profesores de Inmunología, se realizó la evaluación de los reactivos, los cuales fueron puestos en línea en la plataforma Moodle, a través de la herramienta "Retroalimentación" (de manera similar a la que se aplicó en el pilotaje descrito en la sección anterior). Se formaron de manera aleatoria 4 parejas de profesores, y se asignó a cada pareja un total de 77 reactivos, asignados también de manera aleatoria; un profesor participó en la evaluación de los 308 reactivos, considerando su evaluación como criterio de desempate en caso de que no hubiera acuerdo en la evaluación por alguna pareja de profesores. Para medir el grado de acuerdo entre los profesores para cada criterio que contempla el instrumento se utilizó el estadístico kappa de Fleiss[57], que a diferencia de la kappa de Cohen, es utilizado para medir concordancia entre múltiples jueces. El análisis estadístico para los índices kappa se realizó con la versión 2013 de Microsoft Excel®.

3.7.3 Evidencia de validez relacionada con la estructura interna

Para la evidencia de validez del uso del instrumento para la evaluación cualitativa de los reactivos se tomaron en cuenta aspectos relacionados con la estructura interna del mismo. Para esto, se analizó la discriminación de los criterios que forman parte del instrumento, la confiabilidad y la estructura factorial. Para el primero se utilizaron la correlación punto-biserial de Pearson (Rpbis) y una prueba t de Student; para la confiabilidad se utilizó el alfa de Cronbach como un estadístico de consistencia interna; para definir las dimensiones del instrumento se realizó un análisis factorial exploratorio que permitió definir la estructura final. Estos análisis se realizaron con la versión 21 del software SPSS®.

3.8 Evaluación cuantitativa de los reactivos

Para el análisis psicométrico de los reactivos evaluados se utilizó la teoría clásica de los test. El análisis estadístico se realizó con la versión 4.2 del software Iteman®. Los estadísticos utilizados para la evaluación cuantitativa corresponden al índice de dificultad y la correlación punto-biserial de cada uno de los reactivos, así como el alfa de Cronbach para la confiabilidad de la prueba[58].

Al no contar con las bases de datos correspondientes a las respuestas de los estudiantes para los exámenes aplicados durante los años 2012 y 2013, el análisis de las características cualitativas de los reactivos únicamente se realizó para los 200 reactivos aplicados en el año escolar 2014.

Como criterios para clasificar los reactivos de acuerdo con su comportamiento psicométrico, se tomaron en cuenta dos clasificaciones diferentes, una propuesta por Downing[59] y otra propuesta por Haladyna[18] (**Tablas 3.2 y 3.3**).

Tabla 3.2 Clasificación de reactivos de acuerdo con su dificultad y discriminación (Traducido de Downing & Yudkowsky, 2009)

Clasificación del reactivo	Dificultad del reactivo	Discriminación del reactivo (punto-biserial)	Descripción
Nivel I	0.45 a 0.75	0.20 o mayor	Los mejores estadísticos para un reactivo; de ser posible, utilizar la mayoría de reactivos en este nivel.
Nivel II	0.76 a 0.91	0.15 o mayor	Fácil; usar con moderación.
Nivel III	0.24 a 0.44	0.10 o mayor	Difícil; usar con moderación y únicamente si el contenido es esencial (reelaborar si es posible).
Nivel IV	<0.24 o >0.91	Cualquier discriminación	Extremadamente difícil o fácil; no utilizar a menos que el contenido sea esencial.

Tabla 3.3 Clasificación de reactivos de acuerdo con su dificultad y discriminación (Traducido de Haladyna, 2013)

Clasificación del reactivo	Dificultad del reactivo	Discriminación del reactivo (punto biserial)	Descripción
Tipo 1	0.6 a 0.9	>0.15	Reactivo ideal. Dificultad moderada y buena discriminación.
Tipo 2	0.6 a 0.9	<0.15	Mala discriminación.
Tipo 3	Más de 0.9	Cualquiera	Reactivo con alto desempeño; usualmente no discrimina bien.
Tipo 4	<0.6	>0.15	Difícil pero con buena discriminación.
Tipo 5	<0.6	<0.15	Difícil y sin buena discriminación.
Tipo 6	<0.6	<0.15	Igual que el tipo 5, pero uno de los distractores tiene un comportamiento como el tipo 1, lo que usualmente indica un error en la clave.

Con el fin de eliminar aquellos resultados obtenidos por estudiantes que pudieran contestar el examen al azar, se definieron valores límite para establecer un punto de corte que permitiera identificarlos. Para el análisis se partió del supuesto de que los resultados obtenidos en cada uno de los reactivos tiene una distribución Bernoulli (es decir, el resultado obtenido al contestar un reactivo es dicotómico y tiene una probabilidad determinada para tomar un valor X, dada la ocurrencia o no del evento[60]). Los valores para X en esta distribución son de uno en caso de que el estudiante conteste de manera correcta el reactivo y cero en caso de que lo conteste de manera equivocada. A partir de un experimento de Bernoulli[61] con 10 millones de simulaciones, se calculó la probabilidad de que un estudiante obtuviera un determinado puntaje únicamente debido al azar. Sólo aquellos puntajes con una probabilidad menor del 5% de haber podido ser obtenidos por el azar ($p<0.05$) fueron incluidos para análisis psicométrico previamente descrito. Se realizaron los cálculos para exámenes conformados por 60 y 70 reactivos (correspondientes a la cantidad de reactivos que conforman los tres exámenes analizados), cada reactivo con 4 opciones de respuesta (**Anexos 8 y 9**).

Para analizar la relación de las calificaciones obtenidas entre los diferentes tipos de reactivos, se clasificaron los resultados obtenidos por los estudiantes en tres grupos diferentes:

- Calificación global (GL). Corresponde al total de aciertos obtenidos por el estudiante en la prueba completa, la cual incluye el total de reactivos aplicados.
- Calificación con reactivos estándar o sin errores en su elaboración (*standard items*, ST). Corresponde al total de aciertos obtenidos por el estudiante, tomando en cuenta únicamente aquellos reactivos de la prueba que no presentaron ningún error en su elaboración.
- Calificación con reactivos con errores (*flawed items*, FT). Corresponde al total de aciertos obtenidos por el estudiante, tomando en cuenta únicamente los reactivos de la prueba que presentaron uno o más errores en su elaboración.

Se realizó una correlación de las calificaciones obtenidas en cada una de las parejas de grupos de reactivos que forman parte de los exámenes. También se realizó un análisis de regresión lineal para definir el grado de predicción de la calificación final por parte de cada uno de los tipos de reactivo analizados (ST contra FL).

A partir de las medias de los valores de dificultad y de discriminación, se utilizó la prueba de Wilcoxon para comparar estos valores entre los reactivos ST y los reactivos FL, debido a que la diferencia en la cantidad de valores para cada una de las variables no permitió asumir la normalidad en ambas muestras[62]. También se utilizó la correlación de Pearson para estimar la asociación entre el número de errores en el reactivo y la modificación de los parámetros psicométricos.

Finalmente, se realizó la cuantificación de distractores no funcionales, y se calculó el coeficiente de correlación de Pearson entre el número de NFD y la dificultad y la discriminación de los reactivos.

3.9 Aspectos éticos

Para esta investigación se tomó en cuenta y no se violó ningún principio de la Declaración de Helsinki[63], principalmente porque el objeto de medición en el estudio son los propios reactivos, no los estudiantes ni los profesores. La participación de los diferentes evaluadores fue totalmente voluntaria. Este proyecto también se realizó apegándose a los principios del código de ética de la *American Educational Research Association* (AERA), en cuanto a competencia profesional, integridad, responsabilidad profesional, científica y escolar, respeto por los derechos, dignidad y diversidad de las personas, y responsabilidad social[64].

Este proyecto igualmente se apega a los lineamientos establecidos por el H. Consejo Técnico de la Facultad de Medicina sobre la institucionalización del resguardo de los bancos de reactivos para evaluaciones de la Facultad de Medicina[65], asumiendo que los reactivos son información reservada y, por lo tanto, es responsabilidad del jefe del departamento académico, coordinadores de enseñanza y de evaluación garantizar su resguardo, seguridad y evitar que sean difundidos.

4. Resultados

En este apartado se describen los resultados obtenidos a partir de la evaluación cualitativa y cuantitativa de los reactivos, y al final se integran ambos componentes para describir las características de los mismos.

4.1 Evaluación cualitativa de los reactivos

4.1.1 Evidencia de validez relacionada con el contenido.

La revisión de los profesores expertos en el área de evaluación permitió definir el instrumento que aplicaron los profesores de Inmunología para la evaluación de reactivos. Dentro de los aspectos en los que contribuyó la participación de los primeros está la redacción de los indicadores y la unificación de aquellos que resultaban redundantes. Este instrumento estuvo integrado por 22 indicadores diferentes, el primero relacionado con el nivel taxonómico del reactivo y los otros 21 por los criterios que permiten evaluar la calidad del reactivo.

4.1.2 Evidencia de validez relacionada con el proceso de respuesta.

Con el instrumento resultado de la evaluación por expertos se realizó la evaluación de los 308 reactivos seleccionados, de acuerdo con las características mencionadas en el método. Para esta fuente de evidencia de validez se calculó el índice kappa de Fleiss (**Tabla 4.1**).

El índice kappa para el nivel taxonómico de los reactivos fue de 0.189. Para 20 de los 21 criterios de calidad de los reactivos el índice kappa tuvo valores que van desde 0.572 hasta 1. El criterio 20 (“¿Se evita el uso de términos como SIEMPRE, NUNCA, COMPLETAMENTE o ABSOLUTAMENTE?”) se comportó como una constante (con un índice kappa de 1), por lo que no fue incluido en el resto de los análisis. El criterio 11 (“¿Es posible responder la pregunta sin necesidad

de observar las respuestas?") obtuvo un índice kappa negativo (con un valor de -0.23).

Tabla 4.1. Concordancia interjueces en la propuesta de instrumento

Criterio		Índice kappa
	Nivel taxonómico (memoria, comprensión, aplicación).	0.1899
1	¿El reactivo presenta un solo contenido temático?	0.7955
2	¿El reactivo presenta un solo resultado de aprendizaje?	0.7244
3	¿El contenido evaluado está en relación con la especificación del reactivo?	0.9003
4	¿El contenido del reactivo se refiere a una evidencia y no a una opinión?	0.9853
5	¿La semántica utilizada está de acuerdo con el contenido del programa académico?	0.9457
6	¿Las opciones de respuesta se presentan en vertical?	0.9853
7	¿El reactivo cuenta con una gramática, puntuación y ortografía correctas?	0.5906
8	¿La cantidad de texto en el tallo es adecuada para su comprensión?	0.7739
9	¿El tallo del reactivo plantea la idea central?	0.8643
10	¿La pregunta o instrucción se encuentra redactada con claridad?	0.5724
11	¿Es posible responder la pregunta sin necesidad de observar las respuestas?	-0.2318
12	¿El reactivo está expresado en forma positiva (es decir, no incluye palabras como NO o EXCEPTO)?	0.9705
13	¿El reactivo cuenta con tres o cuatro opciones de respuesta?	0.8276
14	¿El reactivo cuenta únicamente con una respuesta correcta?	0.8798
15	¿Las opciones son independientes entre sí?	0.9105
16	¿Las opciones son similares en cuanto a estructura gramatical, contenido y extensión?	0.6086
17	¿Las opciones se expresan de manera afirmativa?	0.9755
18	¿Los distractores son plausibles, es decir, no se descartan por inferencia lógica o sentido común?	0.6558
19	¿Las opciones evitan dar pistas sobre la respuesta correcta?	0.7575
20	¿Se evita el uso de términos como SIEMPRE, NUNCA, COMPLETAMENTE o ABSOLUTAMENTE?	1
21	¿Se evita el uso de las opciones "Todas las anteriores" o "Ninguna de las anteriores"?	0.9902

Los valores de kappa obtenidos para la concordancia entre jueces en la aplicación de los diferentes criterios del instrumento indican que 18 de los 21 ítems analizados obtienen puntajes apropiados (de .060 a 1)[66].

4.1.3 Evidencia de validez relacionada con la estructura interna

La correlación punto-biserial (Rpbis) obtenida entre las diferentes parejas de ítems reflejaron, en general, poca fuerza de asociación (menores de 0.1 y mayores de -0.1) (**Anexo 10**). Sin embargo, hubo cinco correlaciones entre ítems mayores de 0.40 ($p<0.01$) (**Tabla 4.2**)

Tabla 4.2 Correlaciones entre criterios del instrumento

Criterio A	Criterio B	Correlación
1 ¿El reactivo presenta un solo contenido temático?	2 ¿El reactivo presenta un solo resultado de aprendizaje?	0.704 ($p<0.01$)
8 ¿La cantidad de texto en el tallo es adecuada para su comprensión?	10 ¿La pregunta o instrucción se encuentra redactada con claridad?	0.521 ($p<0.01$)
8 ¿La cantidad de texto en el tallo es adecuada para su comprensión?	9 ¿El tallo del reactivo plantea la idea central?	0.497 ($p<0.01$)
14 ¿El reactivo cuenta únicamente con una respuesta correcta?	15 ¿Las opciones son independientes entre sí?	0.424 ($p<0.01$)
9 ¿El tallo del reactivo plantea la idea central?	10 ¿La pregunta o instrucción se encuentra redactada con claridad?	0.417 ($p<0.01$)
7 ¿El reactivo cuenta con una gramática, puntuación y ortografía correctas?	10 ¿La pregunta o instrucción se encuentra redactada con claridad?	0.415 ($p<0.01$)

Las diferentes correlaciones entre los criterios demuestran que el criterio 10 correlaciona de manera significativa directa y moderada con los reactivos 7, 8 y 9; de la misma forma, el reactivo 8 correlaciona con los reactivos 9 y 10. En las correlaciones obtenidas se aprecia además congruencia teórica entre los reactivos.

El análisis de confiabilidad evaluó la consistencia interna del instrumento a través del alfa de Cronbach, con un valor de 0.615 para los 20 elementos cuantificados (como se mencionó previamente, el criterio número 20 del instrumento ya no fue considerado para estos análisis). Al eliminar el criterio 11, la confiabilidad aumentó

a 0.641 (tomando en cuenta el parámetro del análisis “Alfa de Cronbach si se elimina el elemento”) (**Anexo 11**).

La prueba T determinó una p>0.01 para los criterios 4, 6, 12, 17 y 21, por lo que se decidió excluirlos de la versión final del instrumento (**Anexo 12**).

Con los 14 criterios restantes se realizó el análisis factorial exploratorio, en el cual se realizó una rotación oblicua por el método Oblimin, asumiendo que los diferentes factores están correlacionados[30] (**Anexo 13**). La media de adecuación muestral de Kaiser-Meyer-Olkin fue de 0.666 y la prueba de esfericidad de Bartlett arrojó un [χ^2=599.285, r<0.01]. A partir de este análisis exploratorio se lograron identificar cinco factores, cuatro de ellos incluían al menos tres indicadores en su estructura, mientras que el quinto factor sólo tenía un indicador. La varianza explicada del instrumento con cuatro factores fue de 49.979 y con los cinco factores fue de 57.561. El alfa de Cronbach para el instrumento con los 14 ítems finales fue de 0.627. De acuerdo con los elementos que integraron cada factor, se decidió clasificarlas como “Comprensión del reactivo”, “Contenido del reactivo”, “Precisión del reactivo” y “Redacción de opciones de respuesta”. La agrupación de los diferentes criterios a partir del análisis factorial y el instrumento final para la evaluación de reactivos se muestra en la **Tabla 4.3**.

4.1.4 Características cualitativas de los reactivos

La definición del nivel taxonómico del reactivo se verificó a partir del consenso de al menos dos de los jueces, de manera similar a lo reportado por Kibble[66]. Las frecuencias de los niveles taxonómicos de los reactivos analizados se muestran en la **Figura 4.1**. Aquellos reactivos en donde no hubo acuerdo por ningún profesor se incluyen también en esta tabla.

Tabla 4.3 Matriz de estructura de análisis de componentes principales para el instrumento final para la evaluación de reactivos de opción múltiple

Factor	Criterio del instrumento final	Carga	Alfa de Cronbach
Comprensión del reactivo	1. ¿La cantidad de texto en el tallo es adecuada para su comprensión?	.738	0.668 Sig .000
	2. ¿La pregunta o instrucción se encuentra redactada con claridad?	.721	
	3. ¿El reactivo cuenta con una gramática, puntuación y ortografía correctas?	.656	
	4. ¿El tallo del reactivo plantea la idea central?	.655	
Contenido del reactivo	5. ¿El reactivo presenta un solo resultado de aprendizaje?	.865	0.615 Sig .000
	6. ¿El reactivo presenta un solo contenido temático?	.849	
	7. ¿La semántica utilizada está de acuerdo con el contenido del programa académico?	.411	
Precisión del reactivo	8. ¿El reactivo cuenta únicamente con una respuesta correcta?	.814	0.477 Sig .230
	9. ¿Las opciones son independientes entre sí?	.783	
	10. ¿El contenido evaluado está en relación con la especificación del reactivo?	.335	
Redacción de opciones de respuesta	11. ¿Las opciones son similares en cuanto a estructura gramatical, contenido y extensión?	.711	0.357 Sig. 0.44
	12. ¿Las opciones evitan dar pistas sobre la respuesta correcta?	.642	
	13. ¿Los distractores son plausibles, es decir, no se descartan por inferencia lógica o sentido común?	.608	
*	14. ¿El reactivo cuenta con tres o cuatro opciones de respuesta?	.773	-

* - Se consideró como indicador, ya que no hubo más criterios con los cuales conformar un factor.

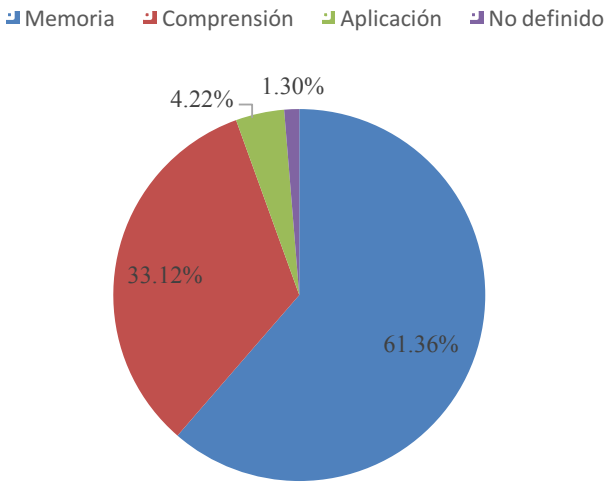


Figura 4.1 Distribución de los niveles taxonómicos entre los 308 reactivos evaluados

Para definir si un reactivo se apegaba o no a los criterios de calidad evaluados en el instrumento previamente descrito, se tomó también el consenso de dos jueces, a partir del hecho de que obtuvieron niveles adecuados del índice kappa para ellos.

Al tomar en cuenta los resultados obtenidos con el instrumento original de 21 indicadores, el número de fallas identificadas en los reactivos va de 0 hasta un máximo de 5, y estos errores ocurrieron en 91 de los 308 reactivos analizados (esto es, tienen una prevalencia del 29.55%), con 58 reactivos identificados con 1 error en su elaboración, 15 reactivos con 2 errores, 11 reactivos con 3 errores, 6 reactivos con 4 errores y 1 reactivo con 5 errores (**Figura 4.2**). Al tomar en cuenta estos datos, se estableció que el 70.45% de los reactivos no presenta errores en su elaboración.

Al realizar la corrección de estos resultados, tomando en cuenta únicamente los 14 indicadores del instrumento final, resultado del proceso de validación previamente descrito, el número total de errores va de 0 a 3, teniendo un total de 59 reactivos con al menos un error en su elaboración (19.15%) y 249 sin errores (80.85%). Del total de reactivos con errores, 37 (12.01%) tienen un error, 13 (4.22%)

tienen dos errores y 9 (2.92%) tienen un total de 3 errores en su elaboración (**Figura 4.3**)

A partir del análisis de los diferentes criterios de calidad considerados en la evaluación de los reactivos, se observa que los errores en la elaboración se distribuyen entre los diferentes indicadores que contempla el instrumento.

Al analizar los resultados obtenidos con el instrumento de 21 indicadores, se observa una prevalencia de apego a las recomendaciones que va desde el 83.8% (correspondiente al indicador "¿Es posible responder la pregunta sin necesidad de observar las respuestas?") hasta el 100% (correspondiente a los indicadores 3, 4, 5, 17, 20 y 21), con una media de cumplimiento de los indicadores de 97.7% (**Figura 4.4**).

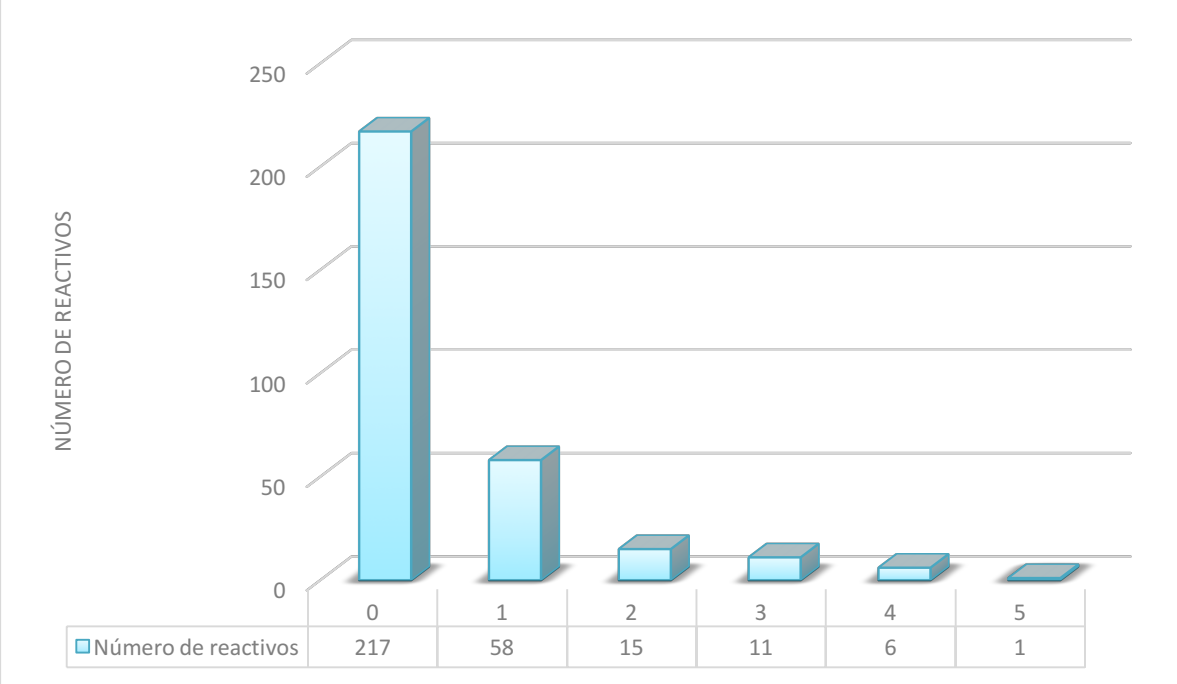


Figura 4.2 Histograma que muestra la distribución de reactivos de acuerdo con el número de errores en su elaboración, a partir del instrumento preliminar de 21 indicadores

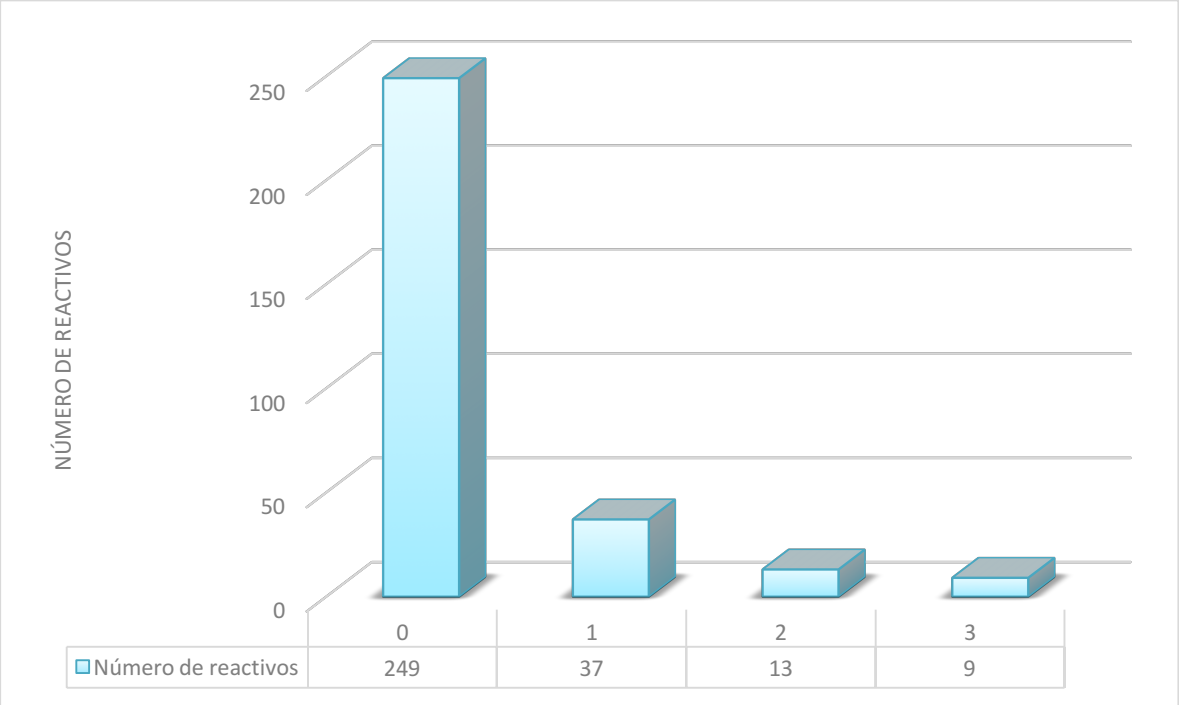


Figura 4.3 Histograma que muestra la distribución de reactivos de acuerdo con el número de errores en su elaboración, a partir del instrumento final de 14 indicadores

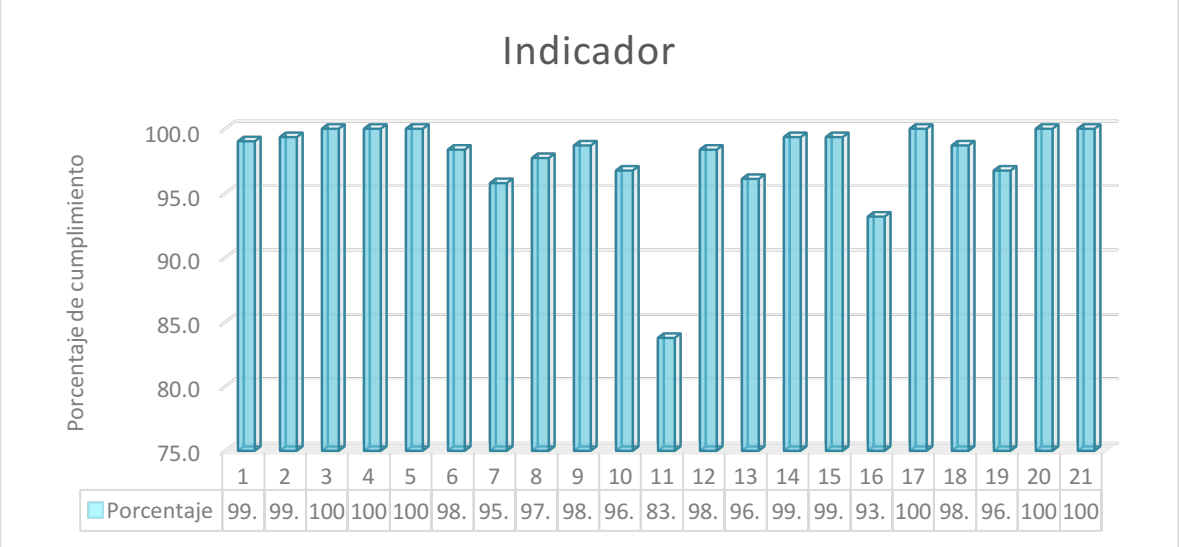


Figura 4.4 Porcentaje de apego a los 21 criterios del instrumento preliminar

Al considerar únicamente los 14 criterios del instrumento final para la evaluación de reactivos, el porcentaje de apego a los criterios adquiere valores que van del 93.2% (correspondiente al indicador "¿Las opciones son similares en cuanto a estructura gramatical, contenido y extensión?") al 100%, con una media de apego del 97.9% (**Figura 4.5**).

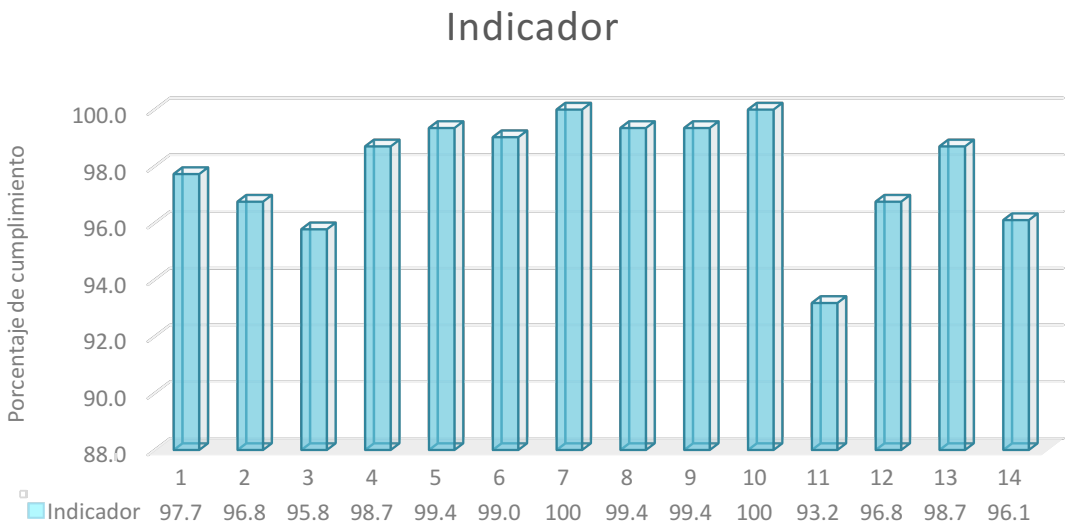


Figura 4.5 Porcentaje de apego a los criterios del instrumento final

Con el fin de realizar la evaluación de los reactivos a partir de sus características cuantitativas y cualitativas, se realizó la medición de las frecuencias de errores en los 200 reactivos en los cuales se hizo el análisis psicométrico (**Figura 4.6**).

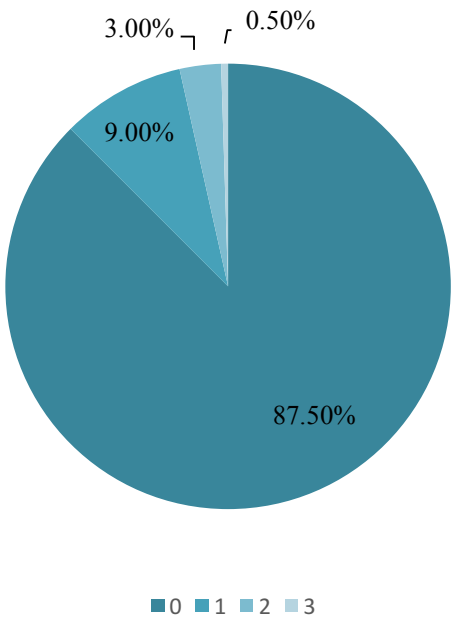


Figura 4.6 Distribución de reactivos de acuerdo con el número de errores en su elaboración, considerando únicamente los 200 reactivos seleccionados para el análisis psicométrico

4.2 Evaluación cuantitativa de los reactivos

4.2.1 Parámetros psicométricos

De acuerdo con los resultados obtenidos en la fase de evaluación cualitativa, se clasificaron los 200 reactivos analizados en tres categorías diferentes: totales (GL, el total de reactivos con los que contaba la prueba), estándar (ST, aquellos que de acuerdo con el instrumento no presentaban ningún error en su elaboración) y deficientes (FL, aquellos que presentaban uno o más errores en su elaboración).

Para precisar los resultados obtenidos en la aplicación de la prueba, se incluyeron únicamente los registros de estudiantes con una baja probabilidad ($p<0.05$) de haber contestado al azar (**Anexos 8 y 9**). Los valores límite para incluir registros fueron de 21 aciertos para el primer examen parcial (P1), conformado por 60 reactivos con 4 opciones de respuesta cada uno, y de 25 aciertos para el

segundo (P2) y tercero (P3), conformados por 70 reactivos con 4 opciones de respuesta cada uno.

Los resultados del análisis psicométrico de los exámenes analizados se muestran en la **Tabla 4.4**. Se realizó el análisis por cada tipo de reactivos (totales, estándar y deficientes, y se observó su comportamiento como exámenes diferentes. Para el análisis, se tomaron en cuenta los valores del índice de dificultad (estimado como la proporción de aciertos entre el número total de reactivos), el de discriminación (a partir de la correlación punto biserial de Pearson), y el alfa de Cronbach como estadístico para calcular la confiabilidad de los exámenes.

Los valores de dificultad de los reactivos analizados tuvieron una media total de 0.596, con un valor mínimo de 0.1 y un máximo de 0.96. Los valores de discriminación tuvieron una media de 0.29, con un mínimo de -0.2 y un máximo de 0.55 (**Figuras 4.7 y 4.8**). Si bien estos valores se muestran de manera conjunta para el total de reactivos (200) del estudio, el análisis psicométrico se realizó tomando en cuenta únicamente el número de sustentantes que presentaron cada prueba y el total de reactivos que integraron el examen en el que fueron aplicados.

El análisis para el índice de dificultad demostró resultados similares en los tres exámenes, tanto en su versión completa (con todos los reactivos), como en los exámenes divididos de acuerdo con la calidad de los reactivos (**Figura 4.9**).

Los índices de discriminación (calculados a través de la correlación punto biserial de Pearson) mostraron también valores similares en los exámenes conformados por el total de reactivos y por los reactivos estándar, mientras que sí existe una disminución en este índice entre los exámenes conformados únicamente por reactivos deficientes (**Figura 4.10**).

Tabla 4.4. Comportamiento psicométrico de los reactivos

Tipo de reactivos	N	Número de reactivos	Dificultad	Discriminación (Rpbis)	Alfa de Cronbach
Primer parcial (P1)					
Porcentaje global (GLPrc)	1071	60	.64	.26	.84
Porcentaje con reactivos estándar (STPrc)		46	.65	.26	.81
Porcentaje con reactivos deficientes (FLPrc)		14	.63	.16	.45
Segundo parcial (P2)					
Porcentaje global (GLPrc)	827	70	.58	.28	.88
Porcentaje con reactivos estándar (STPrc)		63	.59	.27	.86
Porcentaje con reactivos deficientes (FLPrc)		7	.56	.23	.48
Tercer parcial (P3)					
Porcentaje global (GLPrc)	824	70	.57	.33	.91
Porcentaje con reactivos estándar (STPrc)		66	.57	.32	.90
Porcentaje con reactivos deficientes (FLPrc)		4	.58	.24	.42

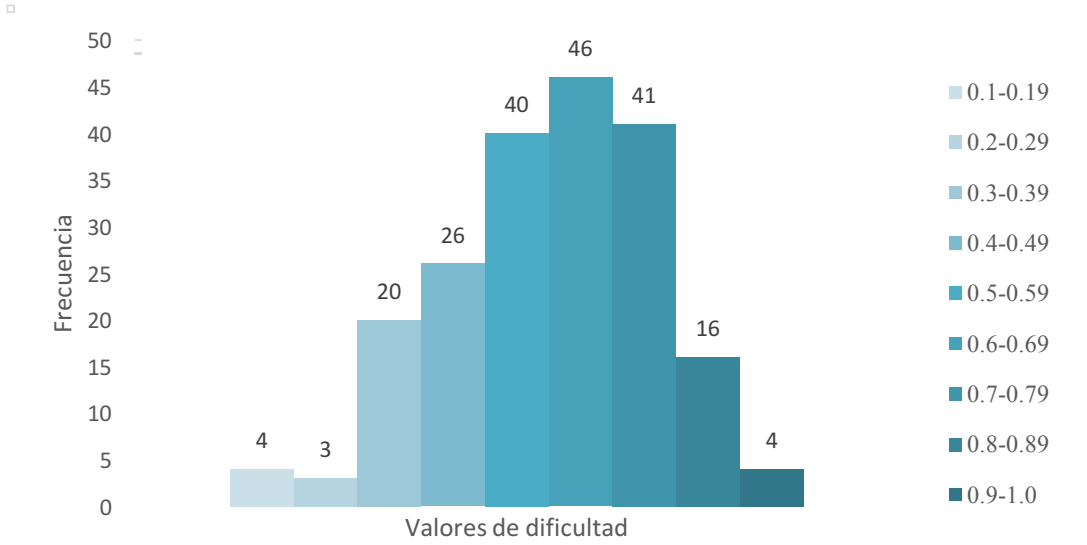


Figura 4.7 Frecuencia de índices de dificultad entre los reactivos evaluados

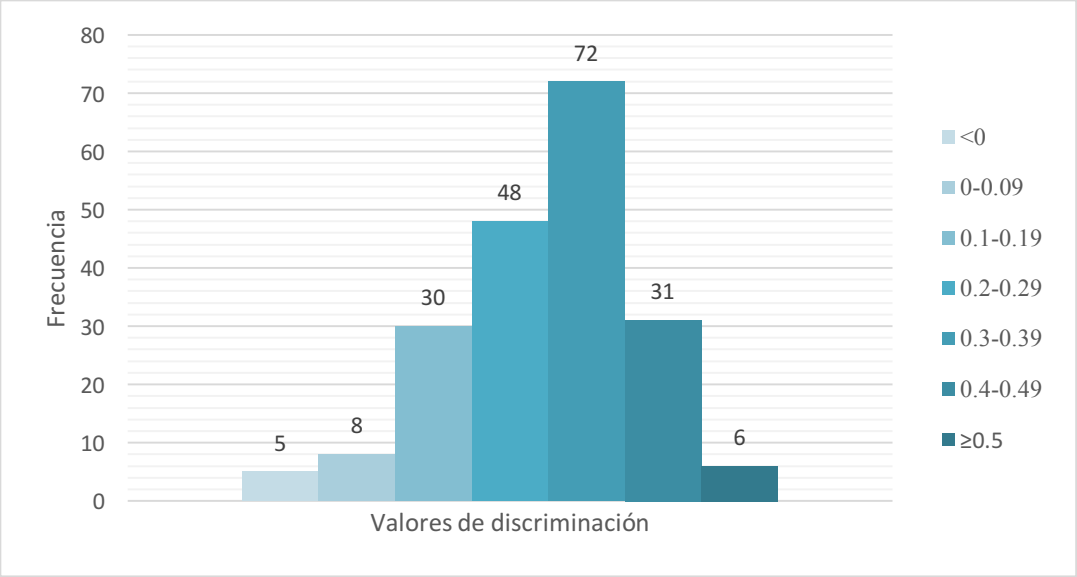


Figura 4.8 Frecuencia de índices de discriminación entre los reactivos evaluados

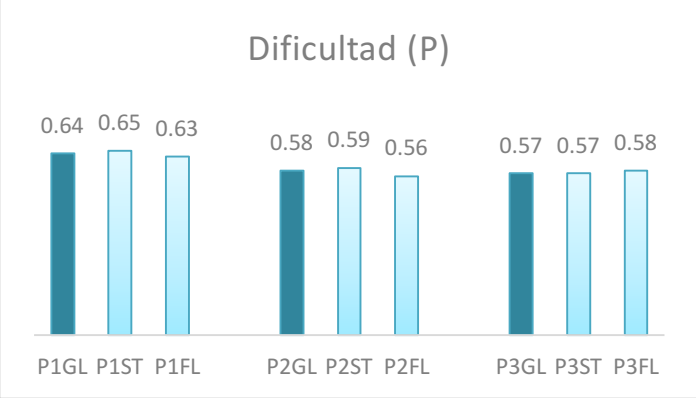


Figura 4.9 Índices de dificultad para cada uno de los exámenes analizados, considerando los diferentes tipos de reactivos (GL, ST y FL)

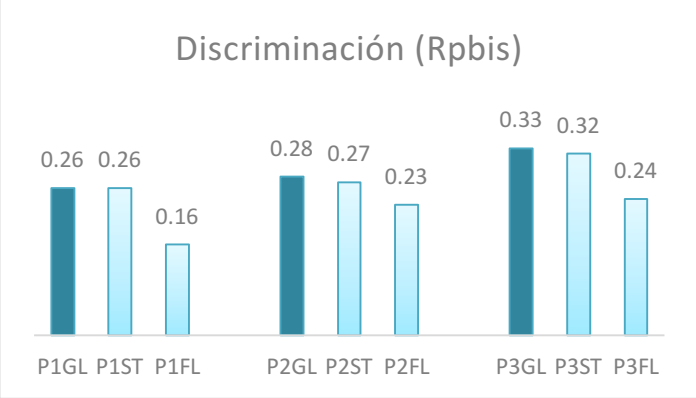


Figura 4.10 Índices de discriminación para cada uno de los exámenes analizados, considerando los diferentes tipos de reactivos (GL, ST y FL)

La confiabilidad de los exámenes, calculada a través del alfa de Cronbach, demostró valores altos del estadístico en las versiones del examen con el total de reactivos y con los reactivos estándar, mientras que sí hubo disminución de estos valores en todos los subgrupos de exámenes integrados por reactivos deficientes (**Figura 4.11**).

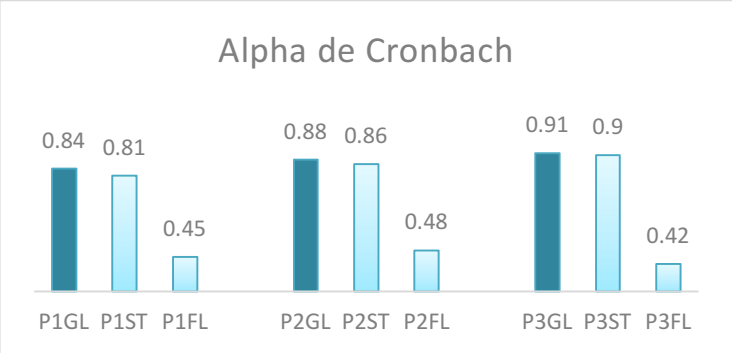


Figura 4.11 Confiabilidad de cada uno de los exámenes analizados, considerando los diferentes tipos de reactivos (GL, ST y FL)

Una vez que se obtuvieron las características psicométricas de los reactivos (**Anexo 14**), se clasificaron de acuerdo con los parámetros obtenidos, con base en las propuestas de Downing[59] y Haladyna[18]. Las **Tablas 4.5 y 4.6** muestran la cantidad de reactivos en cada una de los diferentes niveles o tipos de acuerdo con estas clasificaciones. Debido a que la clasificación de Downing establece rangos definidos para los valores de dificultad y de discriminación correspondientes, un total de 23 reactivos no forman parte de ninguno de los niveles descritos. Para el caso de la clasificación de Haladyna, como se mencionó previamente, los reactivos del tipo 6 son iguales a los del tipo 5, pero uno de los reactivos se comportó como el tipo 1, usualmente bajo el supuesto de que existe un error en la clave para definir la respuesta correcta.

Tabla 4.5 Frecuencia de reactivos de acuerdo con la clasificación de Downing[59]

Clasificación del reactivo	Dificultad del reactivo	Discriminación del reactivo (punto-biserial)	Cantidad de reactivos
Nivel I	0.45 a 0.75	0.20 o mayor	112
Nivel II	0.76 a 0.91	0.15 o mayor	28
Nivel III	0.24 a 0.44	0.10 o mayor	29
Nivel IV	<0.24 o >0.91	Cualquier discriminación	8
Sin clasificación	Cualquiera	Menor del valor requerido para cada nivel	23

Tabla 4.6 Frecuencia de reactivos de acuerdo con la clasificación de Haladyna

Clasificación del reactivo	Dificultad del reactivo	Discriminación del reactivo (punto-biserial)	Cantidad de reactivos
Tipo 1	0.6 a 0.9	>0.15	96
Tipo 2	0.6 a 0.9	<0.15	8
Tipo 3	Más de 0.9	Cualquiera	3
Tipo 4	<0.6	>0.15	81
Tipo 5	<0.6	<0.15	12
Tipo 6	<0.6	<0.15	Ninguno

4.2.2 Distractores no funcionales

Dentro de la muestra total de reactivos evaluados, hubo un total de 800 respuestas, de las cuales, 200 fueron las respuestas correctas y las 600 restantes fueron distractores. De los distractores analizados, un total de 90 (11.25% del total de opciones de respuesta o 15% del total de distractores) fueron catalogados como distractores no funcionales. La distribución de estos distractores se muestra en la **Tabla 4.7**.

Tabla 4.7 Frecuencia de distractores no funcionales en los 200 reactivos analizados

Número de NFD	Cantidad de reactivos	Porcentaje de reactivos
0	134	67%
1	45	22.5%
2	18	9%
3	3	1.5%

La correlación entre el número de NFD en el reactivo y la dificultad fue positiva (0.506, $p<0.01$), mientras que la correlación entre el número de NFD y la discriminación del reactivo fue negativa (-0.253, $p<0.01$) (**Anexo 15**).

4.3 Características cualitativas y cuantitativas de los reactivos

4.3.1 Clasificación de reactivos de acuerdo con sus características

De acuerdo con los resultados obtenidos en las evaluaciones cualitativa y cuantitativa, se clasificaron los reactivos, tomando en cuenta el número de errores en su elaboración y las propuestas de Downing y Haladyna (**Tablas 4.8 y 4.9**).

4.3.2 Calificación del estudiante de acuerdo con el tipo de reactivo

Para identificar si existen variaciones en las calificaciones obtenidas por los estudiantes entre los diferentes tipos de reactivos, estándar (ST) o con errores deficientes debido a errores en su elaboración (FL), con respecto a la calificación global (GL) obtenida en la aplicación del examen, se aplicó una prueba T para muestras relacionadas. Este análisis se realizó para cada uno de los exámenes que formaron parte del estudio (**Tabla 4.10**).

Como se puede observar, tanto en el primer examen (P1) como en el segundo (P2), la media más alta corresponde a la calificación obtenida en los reactivos ST, seguida por la calificación GL, y la media más baja corresponde a los reactivos FL. Esta situación se invierte en el P3, en donde la media más alta corresponde a la calificación con los reactivos FL, seguida de la calificación GL y la media más baja corresponde ahora a los reactivos ST.

Tabla 4.8 Clasificación de los 200 reactivos de acuerdo con la propuesta de Downing (2009) y el número de errores en su elaboración

Número de errores	Nivel I	Nivel II	Nivel III	Nivel IV	Sin clasificar	Total
0	96	24	28	7	20	175
1	12	2	0	1	3	18
2	4	1	1	0	0	6
3	0	1	0	0	0	1
Total:	**112**	**28**	**29**	**8**	**23**	**200**

Tabla 4.9 Clasificación de los 200 reactivos de acuerdo con la propuesta de Haladyna (2013) y el número de errores en su elaboración

Número de errores	Tipo 1	Tipo 2	Tipo 3	Tipo 4	Tipo 5	Total
0	81	8	3	74	9	175
1	10	0	0	5	3	18
2	4	0	0	2	0	6
3	1	0	0	0	0	1
Total:	**96**	**8**	**3**	**81**	**12**	**200**

Tabla 4.10 Estadísticos de las calificaciones obtenidas en los exámenes, clasificadas de acuerdo con los tipos de reactivos que conforman cada tipo de prueba

Tipo de reactivos	N	Media	Puntaje mínimo*	Puntaje máximo	Error estándar de la media
Primer parcial (P1)					
Porcentaje global (GLPrc)	1071	64.343 ±13.93	36.66	96.66	.426
Porcentaje con reactivos estándar (STPrc)		64.809 ±14.68	28.26	100	.449
Porcentaje con reactivos deficientes (FLPrc)		62.811 ±15.95	14.28	100	.487
Segundo parcial (P2)					
Porcentaje global (GLPrc)	827	58.327 ±14.97	37.14	95.71	.521
Porcentaje con reactivos estándar (STPrc)		58.580 ±14.81	33.33	95.23	.515
Porcentaje con reactivos deficientes (FLPrc)		56.054 ±23.27	0	100	.809
Tercer parcial (P3)					
Porcentaje global (GLPrc)	824	56.927 ±17.06	37.14	100	.594
Porcentaje con reactivos estándar (STPrc)		56.851 ±16.95	33.33	100	.591
Porcentaje con reactivos deficientes (FLPrc)		58.191 ±29.13	0	100	1.015

** Este puntaje se obtuvo tras la eliminación de los resultados con alta probabilidad de haber sido obtenidos al azar.*

Las correlaciones observadas para la calificación obtenida en cada pareja de reactivos (GL, ST y FL) por examen se muestra en la **Tabla 4.11**. En estos casos, observamos que para los tres exámenes, la correlación más alta se da entre los reactivos GL y los reactivos ST, mientras que la más baja corresponde siempre a las parejas de reactivos ST y FL ($p<0.05$).

Tabla 4.11 Correlaciones para cada pareja de los tipos de reactivos que conforman los exámenes analizados

Parejas de reactivos comparados	N	Correlación	Significancia
Primer parcial (P1)			
GLPrc y STPrc	1071	.979	.000
GLPrc y FLPrc		.785	.000
STPrc y FLPrc		.642	.000
Segundo parcial (P2)			
GLPrc y STPrc	827	.993	.000
GLPrc y FLPrc		.744	.000
STPrc y FLPrc		.661	.000
Tercer parcial (P3)			
GLPrc y STPrc	824	.997	.000
GLPrc y FLPrc		.678	.000
STPrc y FLPrc		.619	.000

A partir del análisis de regresión lineal se definió la relación que existe entre los puntajes obtenidos en los diferentes tipos de reactivo (ST y FL) y la puntuación global obtenida por el estudiante, de modo tal que los valores obtenidos permitieron identificar el porcentaje en que cada uno de los reactivos permite predecir la calificación final. (**Tabla 4.12**).

Una vez clasificados, se calculó la media de la dificultad y de la discriminación de los diferentes reactivos. Los resultados por cada uno de los exámenes se muestran en las **Tablas 4.13, 4.14 y 4.14**, mientras que los resultados de los 200 reactivos, de manera global, se resumen en la **Tabla 4.16**.

Tabla 4.12 Análisis de regresión lineal para estimar el grado de predicción del puntaje global de cada uno de los tipos de reactivo que conforman el examen

Variable independiente	Variable dependiente	R	R^2
Primer parcial (P1)			
ST	GL	.981	.962
FL		.826	.682
Segundo parcial (P2)			
ST	GL	.994	.988
FL		.776	.601
Tercer parcial (P3)			
ST	GL	.997	.995
FL		.696	.485

Tabla 4.13 Comportamiento psicométrico de los diferentes tipos de reactivos en el primer examen, conformado por 60 reactivos

Reactivos	N	Dificultad	Discriminación
ST	46	0.649	0.264
FL	14	0.628	0.224

Tabla 4.14 Comportamiento psicométrico de los diferentes tipos de reactivos en el segundo examen, conformado por 70 reactivos

Reactivos	N	Dificultad	Discriminación
ST	63	0.586	0.277
FL	7	0.559	0.320

Tabla 4.15 Comportamiento psicométrico de los diferentes tipos de reactivos en el tercer examen, conformado por 70 reactivos

Reactivos	N	Dificultad	Discriminación
ST	66	0.568	0.325
FL	4	0.583	0.375

Tabla 4.16 Comportamiento psicométrico de los diferentes tipos de reactivos tomando en cuenta el total analizado (200 reactivos)

Reactivos	N	Dificultad	Discriminación
ST	175	0.596	0.292
FL	25	0.601	0.275

Al realizar la prueba de Wilcoxon de los rangos con signo de muestras relacionadas se demostró que la mediana de las diferencias entre la dificultad de los reactivos ST y de los reactivos FL es igual a cero (0), sin presentar significancia estadística (p=.085); para la mediana de las diferencias entre la discriminación de ambos tipos de reactivos el valor también fue de cero (0), sin significancia estadística (p=.226) (**Anexo 16**).

Al realizar la correlación entre el número de errores en un reactivo y los valores psicométricos, se demostró que existe una correlación positiva (0.042) entre el número de errores y el índice de dificultad del reactivo, mientras que esta correlación es negativa (-0.10) para la correlación punto biserial. Ninguno de los dos parámetros fue estadísticamente significativo (p=0.556 y 0.884, respectivamente) (**Anexo 17**).

Al establecer la calificación del estudiante tomando en cuenta tanto los reactivos totales (GL) como los reactivos estándar (ST), se observa que hay una mayor proporción de estudiantes aprobados considerando GL que con ST (**Tabla 4.17**), tomando como criterio de aprobación una calificación mayor o igual a 6.0 (60% o más de respuestas correctas en el examen).

Tabla 4.17 Número de estudiantes aprobados considerando la calificación global (GL) y la obtenida con los reactivos estándar (ST)

	Total de sustentantes	Total de aprobados con GL	Total de aprobados con ST	Porcentaje de aprobados con GL	Porcentaje de aprobados con ST
P1	1071	657	650	61.3	60.7
P2	827	375	374	35.0	34.9
P3	824	307	303	28.7	28.3

5. Discusión

Los exámenes con reactivos de opción múltiple son uno de los instrumentos más utilizados para la evaluación de la competencia profesional en medicina[67]. De acuerdo con el concepto actual de validez, es necesario fundamentar a partir de argumentos precisos cada una de las interpretaciones que se realizan a partir de los resultados obtenidos en la evaluación[68]. La literatura actualmente es basta con respecto a las implicaciones que tiene la calidad en la elaboración de reactivos como parte de la validez del proceso de evaluación en donde se utilizan.

Una de las principales amenazas a la validez de una prueba depende de la varianza irrelevante al constructo (CIV), que se refiere al error sistemático en los resultados de la evaluación que depende de variables diferentes al constructo[69]. Dentro de estos factores, los errores en la elaboración de reactivos contribuyen de manera importante a este factor.

En este proyecto se identifican las diferentes evidencias que contribuyen a la validez de la evaluación de reactivos de opción múltiple, para finalmente identificar las características de estos reactivos y las implicaciones que tienen estas características en la calificación de los estudiantes.

5.1 Consideraciones sobre el instrumento para la evaluación de reactivos[70]

El instrumento desarrollado para la evaluación de las características cualitativas de los reactivos se desarrolló tomando en cuenta los aspectos relacionados con las evidencias de validez relacionadas con el contenido, el proceso de respuesta y la estructura interna del mismo, apegándose a los estándares de evaluación internacionales[4] y a las fuentes de evidencia de validez referidas por Messick[29].

La evidencia de validez relacionada con el contenido se refleja a partir de la representatividad del dominio (las características de un reactivo bien elaborado). Al tomar en cuenta las diferentes propuestas de recomendaciones disponibles en la literatura se logró incluir aquellos aspectos considerados a nivel internacional como "deseables" en la elaboración de reactivos de opción múltiple, proceso que fue complementado por la revisión por jueces expertos en el área de evaluación educativa; a partir de las recomendaciones de estos últimos se logró mejorar la propuesta del instrumento con indicadores más entendibles para que resultara más fácilmente aplicable por los evaluadores de los reactivos de Inmunología. La evaluación por jueces representa, por sí misma, una fuente de evidencia de validez[25].

Como parte de la fiabilidad de los resultados, la concordancia entre observadores resulta importante[71]. En este estudio se observó un alto valor en este parámetro (medido a través del estadístico kappa de Fleiss), lo que nos permite inferir que los criterios que forman parte del instrumento para evaluar la calidad de reactivos son entendidos de manera similar por los evaluadores, por lo que resulta adecuado para su aplicación por diferentes jueces. En el caso de este estudio, el criterio 20 ("¿Se evita el uso de términos como SIEMPRE, NUNCA, COMPLETAMENTE o ABSOLUTAMENTE?") es uno de los pocos criterios del instrumento que no están sujetos a la interpretación del evaluador; este comportamiento es similar a otros indicadores que conforman el instrumento, como son los ítems 6, 12, 13, 17 y 21, en los cuales se observó un índice superior a 0.8, lo cual puede ser definido como un "acuerdo casi perfecto"[72]. El resto de los criterios que forman parte del instrumento implican cierta subjetividad al evaluar el reactivo, ya que se requiere de la interpretación del evaluador para criterios que no representan hechos absolutos y que, dependiendo del enfoque o formación del evaluador, pueden llevar a una apreciación diferente del reactivo por los diferentes jueces. A pesar de que en el presente estudios estos indicadores no mostraron una gran variabilidad y que, por lo tanto, no aportan mucha información para poder discriminar los reactivos de acuerdo con su calidad, esto no implica necesariamente

que estas recomendaciones deban ser ignoradas al momento de elaborar reactivos de opción múltiple. Si bien con los fines de este estudio no fueron considerados para formar parte del instrumento final, se deben considerar como "puntos de buena práctica", debido a la evidencia que existe de su importancia en la elaboración de reactivos[20].

Aunque la concordancia para la evaluación de los criterios de calidad fue en general alta, esto no ocurrió con la concordancia para la asignación del nivel cognitivo que evalúan los reactivos. Este resultado ha sido reportado previamente por diversos autores[66,73], quienes refieren que el nivel asignado por los evaluadores depende de diferentes factores, como por ejemplo, el nivel educativo de la persona que se enfrenta al reactivo: la forma de solucionar un problema puede implicar procesos cognitivos diferentes entre varios individuos; por ejemplo, al solucionar un reactivo en donde se solicita asignar un tratamiento para una enfermedad, un estudiante puede utilizar el juicio clínico para solucionarlo (lo cual implica un proceso cognitivo de alto nivel), mientras que otro pueda hacer una asociación memorística para el mismo fin (un proceso cognitivo de bajo nivel). Si bien es controversial la asignación de un nivel taxonómico a un reactivo, se ha reportado que cuando la evaluación se realiza por los profesores que imparten la asignatura y que participan en la elaboración y discusión de los reactivos de un examen, estos valores de concordancia suelen ser mayores[74].

Al analizar las correlaciones entre los diferentes criterios que conforman el instrumento, se pudieron identificar asociaciones importantes entre diferentes parejas de reactivos, como se muestra en la **Tabla 4.2**. El indicador número 10 ("¿La pregunta o instrucción se encuentra redactada con claridad?") tuvo una correlación alta con tres criterios: el 7, que hace referencia a la correcta gramática, puntuación y ortografía del reactivo, el 8, que se refiere a la extensión adecuada del tallo para su comprensión, y el 9, que evalúa si la idea central se encuentra en el tallo; se puede identificar que la presencia de estos últimos tres criterios están indudablemente asociados a que la redacción del tallo sea clara. Para la correlación

entre los criterios 14 y 15, se puede considerar que el hecho de que las opciones de respuesta sean independientes entre sí (esto eso, que una no incluye parcial o totalmente a otra) hace menos probable que el reactivo pueda tener más de una respuesta correcta. La última correlación importante se dio entre los indicadores 1 y 2, en donde resulta evidente que exista una asociación entre evaluar un único contenido temático y un resultado de aprendizaje, incluso la elaboración de un reactivo que evalúe más de un contenido temático o resultado de aprendizaje tiene una mayor dificultad.

El análisis con la prueba T identificó cinco criterios (4, 6, 12, 17 y 21) que no permitían discriminar de manera adecuada las características de los reactivos evaluados; debido a que no aportaban información suficiente fueron eliminados de la versión final del instrumento.

El criterio 11 ("¿Es posible responder la pregunta sin necesidad de observar las respuestas?") incurrió en varias situaciones que llevaron a la decisión de también eliminarlo de la versión final del instrumento. Estos aspectos son los siguientes:

- Este criterio es el único que no está incluido dentro de las recomendaciones originales de Haladyna[20].
- La concordancia entre jueces (medida a través del índice kappa) tuvo un valor negativo, lo que significa que el acuerdo entre los diferentes evaluadores es menor incluso que el esperado por el azar; esto demuestra que el reactivo es poco confiable para su aplicación, ya que está sujeto a la interpretación de la persona que evalúa el reactivo.
- La confiabilidad del instrumento aumenta al eliminar el criterio del instrumento (de 0.615 aumentó a 0.641), fue el único elemento de los 21 criterios en donde este aumento en la confiabilidad ocurrió.
- El reactivo no aportó información de contenido en la estructura de la prueba; a pesar de que su peso factorial fue alto, no se pudo asociar con ninguno de los factores que conformaron la versión final del instrumento.

El análisis factorial identificó cuatro factores diferentes y un indicador aislado. Los primeros cuatro factores logran explicar casi el 50% de la varianza total. El primer factor se integró por los criterios 8, 10, 9 y 7, que como se mencionó previamente, tuvieron una alta correlación; a este factor se le denominó "comprensión del reactivo" y hace referencia a la adecuada redacción del reactivo para permitir su comprensión, esto incluye aspectos formales y de estilo. El segundo factor conformó al componente "contenido del reactivo", estuvo integrado por los criterios 2, 1 y 5, y se refiere a la pertinencia del contenido del reactivo con lo propuesto en el programa académico de la asignatura. El tercer factor estuvo constituido por los criterios 14, 15 y 3, y se nombró "precisión del reactivo", haciendo referencia a la relación del reactivo con una sola especificación y una única respuesta correcta. Por último, el cuarto factor estuvo integrado por los criterios 16, 19 y 18 y fue nombrado "redacción de opciones de respuesta", lo cual se refiere a la adecuada redacción de estos componentes del reactivo.

De acuerdo con las fuentes de evidencia de validez que propone Downing[25], las diferentes propuestas de recomendaciones para elaborar o evaluar reactivos de opción múltiple que se encuentran en la literatura cuentan fundamentalmente con evidencias de validez relacionadas con el contenido. Este instrumento cuenta con tres fuentes diferentes de evidencia de validez (relacionadas con el contenido, con el proceso de respuesta y con la estructura interna del instrumento); lo anterior es de vital importancia para apoyar su uso para la evaluación de la calidad de los reactivos de opción múltiple, que es para lo cual fue diseñado.

5.2 Características cualitativas de los reactivos

La aplicación del instrumento para evaluar la calidad de los reactivos permitió identificar el nivel taxonómico del tipo de conocimiento que evalúan, así como el grado de apego a las recomendaciones en su elaboración.

Más de la mitad (61.36%) de los reactivos evaluados en este estudio fueron clasificados en el nivel de "Memoria", aproximadamente un tercio (33.12%) en el nivel de "Comprensión" y sólo el 4.22% en el nivel de "Aplicación". Otros autores han encontrado una prevalencia similar al estudiar reactivos en diferentes escuelas[66,73,75], habiendo siempre un predominio de reactivos que evalúan el primer nivel de la taxonomía de Bloom. Se ha propuesto que es ideal realizar preguntas que evalúen niveles cognitivos superiores en lugar de únicamente el recuerdo de hechos aislados[76]. Para el caso de los reactivos de este estudio, hay un predominio del segundo tipo y una proporción muy pequeña de aquellos que miden procesos cognitivos más complejos. Es importante tomar en cuenta que estos reactivos fueron elaborados antes de que los profesores recibieran la capacitación sobre las características ideales de los reactivos, por lo que es probable que no fuera prevalente la idea de la importancia de esta situación entre el equipo docente de la asignatura.

Es frecuente encontrar en la literatura una alta tasa de desapego a las recomendaciones para elaborar reactivos de opción múltiple[77]. En nuestro estudio, la prevalencia no fue tan alta comparada con otros estudios[52,53,54] (29.55% con la primera propuesta de instrumento y 19.15% con la versión final del mismo). El error que más frecuentemente se presentó corresponde al indicador 11 de la primera propuesta ("¿Es posible responder la pregunta sin necesidad de observar las opciones de respuesta"?), pero como se mencionó previamente, este indicador no se tomó en cuenta para la versión final debido a la poca información que aportaba al instrumento de manera global; lo anterior puede deberse a la interpretación del mismo, ya que de acuerdo con la definición propia de los reactivos de selección, es necesario identificar la respuesta correcta entre una serie de respuestas que se muestran al sustentante[15]. El otro criterio que tuvo un menor porcentaje de apego estuvo relacionado con la homogeneidad en la redacción de las opciones de respuesta, lo que de algún modo puede estar relacionado con la dificultad que implica para los elaboradores de reactivos plantear distractores plausibles.

Cuando un estudiante responde un reactivo de manera correcta, existe la posibilidad de que haya adivinado la respuesta sin realmente saberla[78]. Los errores en la elaboración de reactivos contribuyen a la varianza irrelevante al constructo (CIV) por diversos factores y están asociados a la adivinación[79]. Uno de los más importantes es referido como "*test-wiseness*", y se refiere a la capacidad que tiene un individuo de utilizar las características y formato de una prueba para aumentar el puntaje obtenido; este factor es independiente del conocimiento que tiene el estudiante sobre el constructo evaluado[80]. Otro factor importante es denominado "*informed guessing*" o "*educated guessing*", que se refiere uso de conocimiento parcial para eliminar las respuestas incorrectas y aumentar la probabilidad de llegar a la respuesta correcta[81]. Estos factores, entre otros, están presentes al momento de contestar exámenes de opción múltiple[82], por lo que deben ser considerados al momento de realizar interpretaciones sobre los resultados obtenidos en la prueba.

5.3 Características cuantitativas de los reactivos

El análisis de los reactivos del estudio se fundamentó en la teoría clásica de los test (TCT). Existen diferentes limitaciones para el uso de este modelo para el análisis de reactivos[39,41,83]; por ejemplo, no existe una definición precisa entre la puntuación verdadera y el significado o relación que tiene sobre el constructo medido, este último es sólo uno de varios componentes que se ven reflejados en la puntuación verdadera[79]; en la TCT se asume que los grupos en los cuales se establecieron los parámetros de distintas pruebas son equiparables, lo cual no se puede garantizar en la práctica; finalmente, los parámetros que corresponden al individuo (calificación verdadera) y a los reactivos (dificultad y discriminación) son dependientes de la prueba y de la muestra de examinados, respectivamente. Esto puede limitar el desarrollo de pruebas y de análisis más sofisticados.

Vale la pena mencionar que si bien actualmente se está optando por utilizar modelos más sofisticados para este análisis, como puede ser la teoría de respuesta

al ítem (TRI), existen varios argumentos por los cuales el modelo utilizado en este estudio resulta adecuado:

- Los supuestos que se tienen que cumplir en la TRI son más difíciles de cumplir (unidimensionalidad e independencia local de los ítems[84]), mientras que los supuestos de la TCT son más fáciles de cumplir con los datos reales de la prueba.
- El análisis estadístico para la TCT es menos complejo y requiere de software menos especializado
- La estimación de los parámetros del modelo es más sencilla (utilizando los índices P y Rpbis para dificultad y discriminación, respectivamente)
- Se requiere una menor muestra para el análisis (aunque no fue el caso de este proyecto)
- El análisis no requiere de una bondad de ajuste tan precisa para asegurar que los datos se ajusten al modelo.
- Es más sencillo comunicar los resultados a un público no experto en el área de psicometría.
- Con muestras suficientemente grandes, los parámetros obtenidos son similares a los que se obtienen con la TRI[85].
- Más de 60 años de aplicación de la TCT con investigación y resultados apoyan su uso en el análisis de ítems.

Aunque no se encuentra documentada, una práctica relativamente frecuente en la Facultad de Medicina es que los estudiantes deciden dar prioridad a alguna asignatura sobre otras, debido a que si la calificación obtenida en algún examen parcial departamental es reprobatoria, tienen que presentar de manera obligatoria examen ordinario, independientemente de la calificación obtenida en el resto de los exámenes parciales[86]; ante esta situación, el estudiante se presenta al examen parcial con el fin de tener registrada su asistencia al mismo y contesta de manera aleatoria y en ocasiones incompleta el examen. Debido a esto, algunos de los puntajes más bajos en la prueba corresponden a estos estudiantes, y no necesariamente implica un bajo aprendizaje o dominio de los conocimientos

evaluados. Si bien esta práctica es poco habitual ante exámenes de medianas o altas consecuencias[79], fue importante eliminar los registros de estudiantes que tuvieran una alta probabilidad de haber realizado esta acción, con el fin de disminuir el error de medición asociado a estos puntajes, que podría verse reflejado en una disminución en la confiabilidad de la prueba[87] y un aumento en el índice de dificultad. (**Anexo 18**).

En la distribución de los índices de dificultad y discriminación se observa una mayor frecuencia de reactivos con dificultad intermedia (0.45 a 0.75, de acuerdo con Downing), así como también una mayor proporción de reactivos con una discriminación adecuada (con un valor de 0.2 o mayor) (**Tabla 4.5**). La recomendación de Downing es que la mayoría de los reactivos que conforman un examen deben de encontrarse dentro de este rango[59], reflejando así un comportamiento psicométrico adecuado en el 56% de los reactivos analizados, este porcentaje aumenta a 70% cuando se incluyen los reactivos del nivel II, que a pesar de ser más fáciles, muestran valores aceptables de discriminación. El 11.5% de los reactivos no pudieron clasificarse en esta propuesta, debido a que el índice de discriminación era menor que el ideal para el índice de dificultad correspondiente, por lo cual son reactivos que no deberían considerarse como parte del banco de reactivos de la asignatura, a reserva de que pudieran modificarse para mejorar su desempeño psicométrico. Al comparar el desempeño psicométrico con las recomendaciones de Haladyna[18] (**Tabla 4.6**), se observa que menos del 50% se apegan al tipo de reactivo que propone como ideal, aunque las consideraciones de Haladyna para este tipo de reactivo son más restrictivas que aquellas propuestas por Downing en cuanto al nivel de dificultad; tomando en cuenta que Haladyna recomienda utilizar solamente los reactivos ubicados en los niveles 1, 3 y 4, el porcentaje de reactivos que forma parte de estas clasificaciones asciende a 90%.

Los valores de confiabilidad para cada uno de los exámenes fueron adecuados, ya que para un examen con ROM, se espera un valor de confiabilidad igual o mayor a 0.8[88]. La consistencia interna es una evidencia de validez relacionada con la

estructura interna de la prueba[89] y se ha demostrado que el uso de ROM en una prueba se asocia a valores de confiabilidad altos, a diferencia de otras pruebas que requieren de calificación manual[90]. Como se mencionó previamente, esta es una de las ventajas del uso de este tipo de instrumento en la evaluación del aprendizaje.

Los distractores que no son elegidos por los sustentantes es muy probable que sean poco plausibles, por lo que no aportan información y deben de ser eliminados[18]. En este estudio, un 33% de los reactivos tuvo al menos un distractor no funcional (NFD), aunque en total únicamente el 15% de todos los distractores analizados fueron clasificados como NFD. En un estudio similar realizado en la misma Facultad de Medicina, se encontró una prevalencia de reactivos con al menos un NFD del 95.7%, y un 53.1% de NFD entre los distractores analizados[91]. Diferentes estudios han demostrado el papel que juegan los distractores en el comportamiento psicométrico de los reactivos. Si bien una de las recomendaciones de Haladyna[20] dice que el reactivo debe de contar con tantas opciones como sea posible, siempre y cuando los distractores sean plausibles, se ha demostrado que el utilizar únicamente tres opciones de respuesta es más sencillo y optimiza el tiempo de respuesta de cada reactivo, sin afectar los valores de dificultad y discriminación e incluso mejorando la confiabilidad de la prueba[92]. Se ha demostrado que los distractores que tienen un índice de dificultad similar (es decir, que son elegidos por proporciones similares de estudiantes en una prueba) aportan más información sobre la habilidad de los sustentantes[93]; en aquellos casos en que un reactivo tiene NFD, la proporción de estudiantes que elige las diferentes opciones de respuesta suele ser muy diferente entre sí; existe evidencia incluso de que a mayor número de NFD en un reactivo, el índice de dificultad es mayor, lo que podría disminuir su capacidad de discriminación[94]. Utilizar reactivos con tres opciones de respuesta disminuye el costo de elaboración y el tiempo de aplicación, lo que permite incluir más reactivos en la prueba[95], aumentando de esta manera la confiabilidad (una evidencia de validez relacionada con la estructura interna de la prueba) y la cantidad de temas o subtemas evaluados (lo cual representa una evidencia de validez relacionada con el contenido).

5.4 Evaluación de los reactivos de opción múltiple

Uno de los pasos más importantes durante el proceso de evaluación y de desarrollo de pruebas (como los exámenes departamentales objeto de este estudio) es la integración del banco de reactivos. Este proceso se debe realizar una vez que los reactivos han sido elaborados, editados, revisados, probados y administrados, reservando en este banco solamente aquellos que tienen las mejores características para su uso potencial en situaciones posteriores[96].

A partir de la clasificación de los reactivos de acuerdo con sus características cualitativas y cuantitativas (**Tablas 4.8 y 4.9**) podemos observar que en este estudio la cantidad de reactivos que se encuentra en el nivel ideal (de acuerdo con Downing) y que no poseen ningún error en su elaboración corresponden a un total de 96 reactivos (48%), en el segundo nivel recomendado se encuentran 24 (12%), lo cual nos dice que sólo el 60% de los reactivos analizados poseen las características adecuadas para ser parte del banco de reactivos de la asignatura; el 40% restante posee alguna característica (ya sea un error en su elaboración o parámetros psicométricos no óptimos) que debe llevar a su revisión y modificación para poder integrarlo en el banco. Al tomar en cuenta la clasificación de Haladyna, el porcentaje de reactivos en el nivel ideal disminuye a un 40.5%, al tomar en cuenta los otros tipos de reactivo que este autor considera adecuados, el porcentaje se eleva a 79%, lo que implica que el 31% restante tendría que ser revisado y modificado para considerar incluirlo en el banco de reactivos.

A partir del cálculo de las medias en la calificación obtenida en los estudiantes de acuerdo con el tipo de reactivo que las conformaba, ya fueran reactivos sin errores (ST), reactivos con errores en su elaboración (FL) o el total de reactivos que conformaban la prueba (GL), se observa que la dificultad de los exámenes se modifica de acuerdo con el tipo de examen; en los dos primeros exámenes, el índice de dificultad es mayor tanto en los exámenes ST como GL, lo que concuerda con lo reportado por Pate sobre cómo los errores en la elaboración pueden hacer que un

examen sea más difícil[53]. Las correlaciones entre reactivos de una misma prueba muestran sus valores más altos entre reactivos GL y ST, mientras que al comparar cualquier tipo de reactivo con los reactivos FL estos valores disminuían considerablemente (de cualquier modo, fueron mayores que lo encontrado en otro estudio por Downing[54]). Por último, en el análisis de regresión lineal para medir el grado de predicción de la GL a partir de los resultados obtenidos en los reactivos ST y de los FL, se observa que el valor es mayor para los ST (muy cercano al valor máximo de 1) que para los FL. Si bien esto puede ser favorecido por la mayor cantidad de reactivos presentes en el grupo ST, no podemos hacer de lado la importancia de cómo la calificación final obtenida por los estudiantes depende en gran manera de los reactivos que se apegan a las recomendaciones de la literatura, y puede ser mejor estimada a partir de ellos.

A partir de estos datos podemos observar también que la dificultad media de los reactivos ST es menor (debido a que tienen un índice P mayor) que en los reactivos FL; en el caso del índice de discriminación, podemos observar que los reactivos ST tienen un valor mayor que los reactivos FL y, por tanto, pueden identificar de mejor manera a los estudiantes de alto y de bajo desempeño. A pesar de que esta diferencia no es estadísticamente significativa, se han obtenido resultados similares en otros estudios[53,54].

La calificación de aprobación en los exámenes departamentales aplicados en la Facultad es con cualquier valor mayor o igual a 6.0, y estos valores se encuentran muy cerca del valor límite, por lo que son los estudiantes que se encuentran más cerca de este valor los que podrían ver afectada su calificación debido a la presencia de reactivos con errores en el examen, ya que esta diferencia en la media de 0.02 es mayor al valor que tiene cada reactivo dentro de la prueba. Downing[54] demostró también que la presencia de errores en la elaboración de reactivos puede tener una repercusión en la acreditación del estudiante, y con los valores obtenidos en este estudio podemos inferir que la presencia de reactivos mal elaborados pueden modificar la decisión de acreditar o no a un estudiante en el examen. Debido a los

valores obtenidos al hacer la correlación entre el número de errores y los parámetros psicométricos, observamos que prácticamente no hay correlación entre estos y la dificultad del reactivo (0.042), pero si hay una correlación negativa (aunque no fue estadísticamente significativa) con la discriminación, esto último podría sugerir que al haber mayor número de errores en el reactivo, su capacidad para identificar a los estudiantes de acuerdo con su nivel de desempeño es menor. En este estudio se demostró que la cantidad de estudiantes que aprueba un examen es diferente si se consideran únicamente los reactivos bien elaborados que cuando se incluyen también los reactivos que tienen errores en su elaboración. Al igual que en el estudio de Downing[54], una proporción de los sustentantes del examen ve modificado el resultado obtenido en el examen e incluso la decisión con respecto a su acreditación al incluir (y considerar para la calificación final) los reactivos mal elaborados, lo cual puede repercutir en la trayectoria escolar del estudiante y tener efectos negativos en el aprendizaje su , así como llevar al evaluador a tomar decisiones equivocadas (a partir de falsos positivos o falsos negativos) durante el proceso educativo. Este último punto también es una fuente de evidencia de validez relacionada con las consecuencias de la evaluación[29], dentro de la propuesta de Messick sobre las diferentes fuentes de evidencia de validez, las cuales deben de tomarse también en cuenta durante el proceso.

5.5 Limitaciones del estudio

Dentro de las limitaciones del estudio se encuentran los siguientes:

- El uso de la teoría clásica de los test parte de supuestos menos sólidos que otros modelos psicométricos, pero anteriormente ya se justificó el uso de este modelo para el estudio.
- Los exámenes analizados corresponden únicamente a una asignatura de la licenciatura de Médico Cirujano, aplicados en sólo una generación de estudiantes. Es necesario dar seguimiento al comportamiento de los reactivos en poblaciones diferentes para aproximarnos más a su valor real y poder realizar inferencias con mayor evidencia de validez.

- Algunas recomendaciones para la elaboración de reactivos no fueron incluidas en la versión final del instrumento y no fueron consideradas para la asociación final de las características cualitativas y cuantitativas de los reactivos. Como se ha referido en diversos estudios[20,77], aún falta evidencia para justificar el apego a ciertas recomendaciones, ya que muchas de ellas cuentan únicamente con el respaldo empírico de expertos en la elaboración de reactivos.

6. Conclusiones

Uno de los principales retos en educación médica es la profesionalización en el área de evaluación por parte de los médicos y académicos del área de la salud[97]. Como parte fundamental de la formación de un médico, la enseñanza y la evaluación de la inmunología representan también un gran reto, ya que al ser una ciencia en desarrollo cuyos contenidos sufren un rápido recambio, los conocimientos actuales pueden resultar medianamente ciertos o incluso totalmente falsos con el paso de los años[98].

La presencia de cualquier error en la elaboración de reactivos debe ser razón suficiente para revisarlo y modificarlo, ya que algunas violaciones a estas recomendaciones pueden modificar el comportamiento psicométrico del reactivo[53]. Uno de los problemas con el uso de ROM es su falta de "validez de apariencia" por parte de los estudiantes, la cual puede ser definida como el grado en que el instrumento aparenta medir la variable deseada[27]; en el caso debido a que no refleja de manera precisa el ejercicio clínico del médico[99]. Si bien la validez de apariencia no forma parte de las evidencias de validez propuestas en la literatura actual, sí puede ser considerada como una característica deseable del proceso de evaluación[26], ya que favorece la aceptación del proceso entre la comunidad educativa.

El objetivo de la licenciatura de Médico Cirujano de la Facultad de Medicina es "formar médicos capaces y competentes para ejercer la medicina general de calidad en ambientes complejos y cambiantes"[47], y sólo una evaluación con la suficiente evidencia de validez nos permitirá garantizar que los estudiantes de la Facultad se conviertan en los médicos que el país necesita.

7. Referencias

1. Wass V, Van Der Vleuten C, Shatzer J, Jones R. Assessment of clinical competence. Lancet. 2001;357(9260):945–9.
2. DLE: evaluación - Diccionario de la lengua española - Edición del Tricentenario [Internet]. [cited 2018 Feb 5]. Available from: http://dle.rae.es/?id=H8JsfPe
3. Scriven M. Evaluation thesaurus [Internet]. 3rd. Library of Congress. California: Library of Congress; 1981. Available from: http://linkinghub.elsevier.com/retrieve/pii/002074899290035F
4. American Educational Research, Association, American Psychological Association, National Council on Measurement in Education. The Standards for Educational and Psychological Testing. American Educational Research Association, editor. Washington, D.C.; 2014.
5. Dirección General de Evaluación Educativa UNAM. Glosario Básico de Términos de Evaluación Educativa [Internet]. [cited 2015 Mar 24]. Available from: http://www.evaluacion.unam.mx/glosario.htm
6. OECD. Synergies for Better Learning [Internet]. OECD Reviews of Evaluation and Assessment in Education. Paris: OECD Publishing; 2013 Apr. Available from: http://www.oecd-ilibrary.org/education/synergies-for-better-learning/student-assessment-putting-the-learner-at-the-centre_9789264190658-7-en
7. Harlen W, Deakin Crick R. A systematic review of the impact of summative assessment and tests on students' motivation for learning (EPPI-Centre Review, version 1.1). Res Evid Educ Libr. London; 2002;(1).
8. Bernad Mainar JA. Modelo cognitivo de evaluacion educativa : escala de estrategias de aprendizaje contextualizado, ESEAC. Narcea; 2000.
9. Committee Development Assistance. Glossary of Key Terms in Evaluation and Results Based Management [Internet]. Paris; 2010 [cited 2018 Feb 18]. Available from: http://www.oecd.org/development/peer-reviews/2754804.pdf
10. Ben-David MF. AMEE Guide No. 18: Standard setting in student assessment. Med Teach [Internet]. 2000 Jan;22(2):120–30. Available from: http://informahealthcare.com/doi/abs/10.1080/01421590078526
11. Downing SM, Yudkowsky R. Introduction to Assessment in the Health Professions. In: Downing SM, Yudkowsky R, editors. Assessment in Health Professions Education. New York: Routledge; 2009. p. 317.
12. Díaz Barriga Arceo F, Hernández Rojas G. Estrategias docentes para un aprendizaje significativo. Diplomado en Informática para la enseñanza de la medicina. México: McGraw-Hill Interamericana; 2002. p. 465.
13. Miller GE. The assessment of clinical skills-competence-performance.pdf. Acad Med. 1990;65(9):S63–7.

14. Downing SM. Assessment of Knowledge with Written Test Forms. In: Norman G, van der Vleuten C, Newble D, editors. International Handbook of Research in Medical Education. Springer Science+Bussiness Media; 2002. p. 647–72.
15. Downing SM. Written Test. In: Downing SM, Yudkowsky R, editors. Assessment in Health Professions Education. New York: Routledge/Taylor & Francis Group.; 2009. p. 149–84.
16. Audiffred Maldonado, Laura Edith et al. Material de apoyo para el taller de elaboración de reactivos del CENEVAL. México; 2009. Report No.: 3.
17. Zimmaro DM, Ph D. Writing Good Multiple-Choice Exams. 2004;(512).
18. Haladyna TM, Rodriguez MC. Developing and validating test items. Routledge; 2013. 446 p.
19. Lunz ME. Examination Development Guidelines [Internet]. Measurement Research Associates, Inc. [cited 2018 Feb 25]. p. 26. Available from: http://www.measurementresearch.com/media/evalguidelines.pdf
20. Haladyna TM, Downing SM, Rodriguez MC. A Review of Multiple-Choice Item-Writing Guidelines for Classroom Assessment. 2002;15(3):309–34.
21. Moreno Olivos T. Evaluación del Aprendizaje y Para el Aprendizaje: reinventar la evaluación en el aula [Internet]. Metropolitana UA, editor. México; 2016. 320 p. Available from: http://www.casadelibrosabiertos.uam.mx/contenido/contenido/Libroelectronico/Evaluacion_del_aprendizaje_.pdf
22. Amin Z, Chong YS, Khoo HE. Practical Guide to Medical Student Assessment [Internet]. World Scientific Publishing Co. Pte. Ltd.; 2006. Available from: http://ebooks.worldscinet.com/ISBN/9789812773586/9789812773586.html
23. Scully D. Constructing Multiple-Choice İtems to Measure Higher-Order Thinking. Pract Assessment, Res Eval. 2017;22(4):1–13.
24. Haladyna TM, Downing SM. A Taxonomy of Multiple-Choice Item-Writing Rules.pdf. Appl Meas Educ. 1989;2(1):37–50.
25. Downing SM. Validity: on the meaningful interpretation of assessment data. Med Educ. 2003 Sep;37(9):830–7.
26. Downing SM, Haladyna TM. Validity threats: overcoming interference with proposed interpretations of assessment data. Med Educ [Internet]. 2004 Mar [cited 2013 Aug 19];38(3):327–33. Available from: http://doi.wiley.com/10.1046/j.1365-2923.2004.01777.x
27. Hernández Sampieri R, Fernández Collado C, Baptista Lucio P. Metodología de la investigación. 6th ed. México: McGraw-Hill Interamericana; 2014. 1-589 p.
28. Cook DA, Brydges R, Ginsburg S, Hatala R. A contemporary approach to validity arguments: A practical guide to Kane's framework. Med Educ. 2015;49(6):560–75.

29. Messick S. Validity of Psychological Assessment: Validation of Inferences from Persons' Responses and Performancesas Scientific Inquiry into Score Meaning. Princeton; 1994.

30. Kerlinger F. Investigacion Del Comportamiento. 4ta. México: McGraw-Hill; 2002. 810 p.

31. DiDonato-Barnes N, Fives H, Krause ES. Using a Table of Specifications to improve teacher-constructed traditional tests: an experimental design. Assess Educ Princ Policy Pract [Internet]. Routledge; 2014 Jan 2 [cited 2018 Jul 19];21(1):90–108. Available from: http://www.tandfonline.com/doi/abs/10.1080/0969594X.2013.808173

32. Alade OM, Omoruyi IV. Table of Specification and Its Relevance in Educational Development Assessment. Eur J Educ Dev Psychol. 2014;2(1):1–17.

33. Fives H, Didonato-Barnes N. Classroom Test Construction : The Power of a Table of Specifications. Pract Assessment, Res Eval. 2013;18(3):1–7.

34. Moreno R, Martínez RJ. Directrices para la construcción de ítems de elección múltiple. Psicothema. 2004;16(2002):490–7.

35. Krathwohl DR. A Revision of Bloom's Taxonomy: An Overview. Theory Pract. 2002;41(4):212–8.

36. Palmer EJ, Devitt PG. Assessment of higher order cognitive skills in undergraduate education: modified essay or multiple choice questions? Research paper. BMC Med Educ [Internet]. 2007 Jan [cited 2013 Oct 19];7:49. Available from: http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=2148038&tool=pmcentrez&rendertype=abstract

37. Buckwalter JA, Schumacher R, Albright JP, Cooper RR. Use of an Educational Taxonomy For Evaluation of Cognitive Performance. J Med Educ. 1981;56:115–21.

38. McGuire C. A Process Approach to the Construction and Analysis of Medical Examinations. J Med Educ. 1963;38(July):556–63.

39. Muñiz J. Las teorías de los tests: teoría clásica y teoría de respuesta a los ítems. Papeles del Psicólogo. 2010;31(1):57–66.

40. Abad FJ, Olea J, Ponsoda V, García C. Construcción de tests y análisis de ítems. Medición en Ciencias Sociales y de la Salud. Madrid: Editorial Síntesis; 2001.

41. Hambleton RK, Jones RW. Comparison of Classical Test Theory and Item Response Theory and Their Applications to Test Development. 1968;38–47.

42. Vallejo PM. La fiabilidad de los tests y escalas. 2007;

43. Ding L, Beichner R. Approaches to data analysis of multiple-choice questions. Phys Rev Spec Top - Phys Educ Res. 2009;5(2):1–17.

44. Haladyna TM, Downing SM. How Many Options is Enough for a Multiple-

Choice Test Item? Educ Psychol Meas [Internet]. Sage PublicationsSage CA: Thousand Oaks, CA; 1993 Dec 7 [cited 2018 Feb 26];53(4):999–1010. Available from: http://journals.sagepub.com/doi/10.1177/0013164493053004013

45. Plan de Estudios 2010 de la Licenciatura de Médico Cirujano [Internet]. 2010. Available from: http://www.facmed.unam.mx/marco/index.php?dir_ver=16
46. Saldaña Balmori Y. Historia del Departamento de Bioquímica de la Facultad de Medicina, Universidad Nacional Autónoma de México. Mens Bioquim. 2018;42:147–66.
47. Plan Único de Estudios de la Carrera de Médico Cirujano [Internet]. 2009 [cited 2018 Feb 18]. Available from: http://www.facmed.unam.mx/plan/unico/index.pdf
48. H. Consejo Técnico de la Facultad de Medicina de la UNAM. Minuta-CTA 9 [Internet]. México; 2013 [cited 2018 Feb 18]. Available from: http://consejo.facmed.unam.mx/home/aaminutas.html
49. Programa Académico de Inmunología [Internet]. 2013. Available from: http://www.facmed.unam.mx/fm/pa/2010/II_inmunologia.pdf
50. H. Consejo Técnico de la Facultad de Medicina de la UNAM. Minuta-CTA 10 [Internet]. 2013 [cited 2018 Jul 20]. Available from: http://consejo.facmed.unam.mx/home/aaminutas.html
51. H. Consejo Técnico de la Facultad de Medicina de la UNAM. Minuta 112 [Internet]. 2012. [cited 2018 Jul 20]. Available from: http://consejo.facmed.unam.mx/home/aaminutas.html
52. Tarrant M, Knierim A, Hayes SK, Ware J. The frequency of item writing flaws in multiple-choice questions used in high stakes nursing assessments. Nurse Educ Pract. 2006 Dec;6(6):354–63.
53. Pate A, Caldwell DJ. Effects of multiple-choice item-writing guideline utilization on item and student performance. Curr Pharm Teach Learn. Elsevier; 2014 Jan;6(1):130–4.
54. Downing SM. The effects of violating standard item writing principles on tests and students: the consequences of using flawed test items on achievement examinations in medical education. Adv Health Sci Educ Theory Pract [Internet]. 2005 Jan [cited 2015 Apr 27];10(2):133–43. Available from: http://www.ncbi.nlm.nih.gov/pubmed/16078098
55. Case SM, Swanson DB. Cómo construir preguntas de Selección Múltiple para Ciencias Básicas y Ciencias Clínicas. 3a. Philadelphia: National Board of Medical Examiners; 2006. 180 p.
56. Downing SM, Haladyna TM. Manual para el desarrollo de pruebas a gran escala. México, D.F.: Centro Nacional de Evaluación para la Educación Superior; 2012.
57. Fleiss JL. Measuring nominal scale agreement among many raters. Psychol

Bull. 1971;76(5):378–82.

58. Guyer R, Thompson NA. User's Manual for Iteman 4.2.1. Assessment Systems Corporation; 2012.
59. Downing SM. Statistics of Testing. In: Downing SM, Yudkowsky R, editors. Assessment in Health Professions Education. New York: Routledge; 2009. p. 317.
60. Madsen B. Statistics for Non-Statisticians [Internet]. Springer, editor. 2011 [cited 2018 Feb 23]. 160 p. Available from: https://link-springer-com.pbidi.unam.mx:2443/content/pdf/10.1007%2F978-3-642-17656-2.pdf
61. Dodge Y. The Concise Encyclopedia of Statistics [Internet]. Springer, editor. New York, NY: Springer New York; 2008 [cited 2018 Feb 23]. 44-45 p. Available from: http://www.springerlink.com/index/10.1007/978-0-387-32833-1_34
62. Wilcoxon F. Individual Comparisons by Ranking Methods. Biometrics Bull [Internet]. International Biometric Society; 1945 Dec [cited 2018 Feb 27];1(6):80. Available from: http://www.jstor.org/stable/10.2307/3001968?origin=crossref
63. Declaración de Helsinki de la AMM – Principios éticos para las investigaciones médicas en seres humanos – WMA – The World Medical Association [Internet]. Asociación Médica Mundial. 2017 [cited 2018 Feb 17]. Available from: https://www.wma.net/es/policies-post/declaracion-de-helsinki-de-la-amm-principios-eticos-para-las-investigaciones-medicas-en-seres-humanos/
64. American Educational Research Association. AERA Code of Ethics: American Educational Research Association Approved by the AERA Council February 2011. Educ Res [Internet]. 2011;40(3):145–56. Available from: http://edr.sagepub.com/cgi/doi/10.3102/0013189X11410403
65. H. Consejo Técnico de la Facultad de Medicina de la UNAM. Minuta-CMU 8 [Internet]. México; 2013 [cited 2018 Feb 18]. Available from: http://consejo.facmed.unam.mx/home/aaminutas.html
66. López A, Galparsoro DU, Fernández P. Medidas de concordancia : el índice de Kappa. Cad Aten Primaria. 1999;6:169–71.
67. Kibble JD, Johnson T. Are faculty predictions or item taxonomies useful for estimating the outcome of multiple-choice examinations? AJP Adv Physiol Educ. 2011;35(4):396–401.
68. Epstein RM, Hundert EM. Defining and Assessing Professional Competence. JAMA. 2002;287(2):226–35.
69. Kane MT. Validating the Interpretations and Uses of Test Scores. J Educ Meas. 2013;50(1):1–73.
70. Downing SM. Construct-irrelevant variance and flawed test questions: Do multiple-choice item-writing principles make any difference? Acad Med. 2002;77(10 Suppl):S103–4.

71. Rivera Jiménez J, Flores Hernández F, Alpuche Hernández A, Martínez González A. Evaluación de reactivos de opción múltiple en medicina. Evidencia de validez de un instrumento. Investig en Educ Médica. Universidad Nacional Autónoma de México, Facultad de Medicina; 2016;6(21):8–15.

72. Landis JR, Koch GG. The measurement of observer agreement for categorical data. Biometrics. 1977 Mar;33(1):159–74.

73. Cunnington JPW, Norman GR, Blake JM, Dauphinee WD, Blackmore DE. Applying Learning Taxonomies to Test Items: Is a Fact an Artifact? Acad Med. 1996;71(10):31–3.

74. Thompson E, Luxton-Reilly A, Whalley JL, Hu M, Robbins P. Bloom's taxonomy for CS assessment. Conf Res Pract Inf Technol Ser. 2008;78(January):155–61.

75. Baig M, Ali SK, Ali S, Huda N. Evaluation of Multiple Choice and Short Essay Question items in Basic Medical Sciences. Pak J Med Sci. 2014;30(1):3–6.

76. Peitzman SJ, Nieman LZ, Gracely EJ. Comparison of "Fact-Recall" with "Higher-order" Questions in Multiple-choice Examinations as Predictors of Clinical Performance of Medical Students. Acad Med. 1990;65(9):S59–60.

77. Martínez RJ, Moreno R, Martín I, Trigo ME. Evaluation of five guidelines for option development in multiple-choice item-writing. 2009;21:326–30.

78. Karandikar RL. Multiple-Choice Test, Negative Marks and an Alternative. Resonance. 2006;(March):86–93.

79. Jurado-Núñez A, Leenen I. Reflexiones sobre adivinar en preguntas de opción múltiple y cómo afecta el resultado del examen. Investig en Educ Médica. 2016;5(17):55–63.

80. Millman J, Bishop CH, Ebel R. An Analysis of Test-Wiseness. Educ Psychol Meas [Internet]. Sage PublicationsSage CA: Thousand Oaks, CA; 1965 Oct 2 [cited 2018 Feb 28];25(3):707–26. Available from: http://journals.sagepub.com/doi/10.1177/001316446502500304

81. Downing SM. Guessing on selected-response examinations. Med Educ. 2003;37(8):670–1.

82. Jurado-Nuñez AG, Rivera Jiménez J, Leenen I. The Effect of Different Scoring Rules on Responding Multiple-Choice Items. Canada International Conference on Education. Infonomics Society; 2017. p. 162–6.

83. Leenen I. Virtudes y limitaciones de la teoría de respuesta al ítem para la evaluación educativa en las ciencias médicas. Investig en Educ Médica. 2014;3(9):40–55.

84. Abad FJ, Olea J, Ponsoda V, García C. Medición en ciencias sociales y de la salud. México: Síntesis; 2011. 556 p.

85. Burton RF. Multiple-choice and true / false tests : myths and misapprehensions. Assess Eval High Educ. 2005;30(1):65–73.

86. Lineamientos para la evaluación del alumnado en la primera fase de la Licenciatura de Médico Cirujano [Internet]. Gaceta Facultad de Medicina. 2014 [cited 2018 Feb 18]. Available from: http://www.facmed.unam.mx/index.php?id_contenido=0000_pu_ga&id_sec=sep292k14&gac_ano=2014

87. Frary RB, Cross LH, Lowry SR, Taylor P. Random Guessing , Guessing , and Correction Reliability Test Scores of for Multiple-Choice. J Exp Educ. 2013;46(1):11–5.

88. Kehoe J. Basic Item Analysis for Multiple-Choice Tests. Pract Assess Res Eval [Internet]. 1995;4(10):8–11. Available from: http://pareonline.net/getvn.asp?v=4&amp

89. Downing SM. Validity : on the meaningful interpretation of assessment data. 2003;830–7.

90. Bodner GM. Statistical Analysis of Multiple Choice Exams. J Chem Educ. 1980;57(3):188–90.

91. Jurado-Nuñez AG, Flores-Fernandez F, Delgado-Maldonado L, Sommer-Cervantes H, Martínez-González A, Sánchez-Mendiola M. Distractores en preguntas de opción múltiple para estudiantes de Medicina ¿Cuál es su comportamiento en un examen de altas consecuencias ? Investig en Educ Médica. 2013;2(8):202–10.

92. Rodriguez MC. Three Options Are Optimal for Multiple-Choice Items : A Meta-Analysis of 80 Years of Research. :3–13.

93. Revuelta J. Analysis of distractor difficulty in multiple-choice items. Psychometrika. 2004;69(2):217–34.

94. Abdulghani H, Ahmad F, Aldrees A, Khalil M, Ponnamperuma G. The relationship between non-functioning distractors and item difficulty of multiple choice questions: A descriptive analysis. J Heal Spec [Internet]. 2014;2(4):148. Available from: http://www.thejhs.org/text.asp?2014/2/4/148/142784

95. Tarrant M, Ware J, Mohammed AM. An assessment of functioning and non-functioning distractors in multiple-choice questions: A descriptive analysis. BMC Med Educ. 2009;9(1):1–8.

96. Downing SM. Twelve Steps for Effective Test Development. In: Downing SM, Haladyna TM, editors. Handbook of Test Development. New Jersey: Taylor & Francis; 2006. p. 778.

97. Sánchez Mendiola M, Delgado Maldonado L, Flores Hernández F, Leenen I, Martínez González A. Evaluación del aprendizaje. In: Sánchez Mendiola M, Lifshitz Guinzberg A, Vilar Puig P, Martínez González A, Varela Ruiz ME, Graue Wierchers E, editors. Educación Médica Teoría y práctica. México: Elsevier; 2015. p. 89–96.

98. Barrera ORS, Rosa D, Rodríguez JR, Rosa D. La Inmunología en la formación de pregrado de la docencia médica. 2005;19(4).

99. Moss E. Multiple choice questions: their value as an assessment tool. Curr Opin Anaesthesiol [Internet]. 2001 Dec;14(6):661–6. Available from: http://www.ncbi.nlm.nih.gov/pubmed/17019162

8. Anexos

Anexo 1. Recomendaciones para la elaboración de reactivos de opción múltiple. Haladyna, Downing, Rodríguez. A review of multiple-choice item-writing guidelines (2002) (traducción de CENEVAL)[55]

Contenido	Todo reactivo debe de reflejar un contenido específico y una sola conducta mental, como sea requerido en las especificaciones de la prueba
	Basar cada reactivo en contenido importante, evitar contenidos triviales
	Usar material novedoso para evaluar aprendizaje del nivel más alto. No usar el mismo lenguaje que en el libro de texto, para evitar evaluar el recuerdo de palabras y oraciones familiares
	Mantener independiente el contenido de cada reactivo
	Evitar contenido demasiado específico o demasiado general
	Evitar reactivos basados en opiniones
	Evitar reactivos engañosos
	Mantener el vocabulario simple y adecuado a los individuos que serán evaluados
Formato	Usar las versiones de pregunta, respuesta y llenado de los reactivos convencionales de opción múltiple, opción alterna, verdadero-falso, múltiples verdadero-falso, igualación, reactivo y grupo de reactivos dependientes de contexto; pero evitar el formato complejo de opción múltiple
	Formatear los reactivos de manera vertical, no horizontal
Estilo	Revisar y editar los reactivos
	Usar gramática, puntuación, ortografía y mayúsculas de manera correcta
	Minimizar la cantidad de lectura requerida para cada reactivo
Tallo	Asegurarse de que las instrucciones en la base sean muy claras
	Incluir la idea central en la base, no en las opciones
	Evitar adornar el reactivo (evite el exceso de palabras)
	Plantear la base de manera positiva, evitar negativos como NO o EXCEPTO. Si se usan palabras negativas, usarlas de manera cuidadosa y siempre asegurarse que la palabra se inicia con mayúsculas y negritas
Opciones de respuesta	Desarrollar tantas opciones efectivas como sea posible, pero recordar que los resultados de investigación sugiere que tres son adecuadas
	Asegurarse de que únicamente una de esas opciones es la correcta
	Variar la posición de la respuesta correcta de acuerdo al número de opciones. Equilibrar la clave de respuestas dentro de lo posible, de modo que la respuesta correcta aparezca el mismo número de veces en cada posición
	Colocar las opciones en orden lógico o numérico
	Mantener las opciones independientes; las opciones no deben superponerse en significado
	Mantener a las opciones homogéneas en estructura gramatical y contenido
	Igualar la extensión de todas las opciones
	Ninguna de las anteriores debe usarse con cuidado
	Evitar el uso de *Todas las anteriores*
	Plantear las opciones de forma positiva; evitar negativos como no
	Evitar dar claves o pistas de la respuesta correcta, tales como: a. Determinantes específicos como siempre, nunca, completamente y absolutamente b. Asociaciones “que suenan”, opciones idénticas o parecidas a la base c. Inconsistencias gramaticales que dan pistas de la respuesta correcta d. Opciones flagrantemente correctas e. Pares o tríos de opciones que dan pistas de la respuesta correcta f. Opciones flagrantemente absurdas y ridículas
	Hacer que todos los distractores sean verosímiles
	Usar errores típicos de los estudiantes para crear distractores
	Usar humor si es compatible con el maestro y el ambiente de aprendizaje

Anexo 2. Tabla de especificaciones de la asignatura de Inmunología para el año escolar 2014

Universidad Nacional Autónoma de México
Facultad de Medicina
Departamento de
Tabla de Especificaciones de Inmunología – Año escolar 2014

Unid	Contenido (Tema)	Subtema	Objetivo temático	Ponderación (sesiones-horas)	Subtotal de reactivos	Nivel de desempeño		
						Conocimiento	Comprensión	Aplicación
1	Generalidades	Definiciones	• Mencionar el desarrollo histórico de esta disciplina. • Definir qué es inmunidad. • Describir los conceptos de lo propio y no propio. • Describir en qué consiste la inmunidad y las diferencias entre inmunidad natural y adquirida, así como inmunidad pasiva y activa. • Describir las características básicas de la respuesta inmune • Comparar las características de la respuesta inmune innata con la adaptativa. • Describir los conceptos de lo que es un antígeno, un anticuerpo y un inmunógeno. • Explicar la proyección y relevancia que la inmunología ha adquirido, en especial en el marco de la medicina.	1 – 2	3	3	-	-
		Órganos linfoides y células del sistema inmune	• Definir los conceptos de órgano linfoide primario y secundario. • Señalar la importancia de: hígado fetal, médula ósea, timo, bazo, ganglios linfáticos y tejidos linfoides no encapsulados asociados a mucosas. • Analizar y relacionar la estructura y función de los órganos linfoides, ya sean primarios o secundarios. • Explicar el concepto de anatomía del sistema inmune en relación con otros aparatos y sistemas. • Describir qué son las células madre pluripotenciales, células progenitoras mieloides y células progenitoras linfoides. • Describir las etapas de diferenciación y maduración de estas poblaciones celulares a través del reconocimiento de la expresión de moléculas de superficie. • Describir los procesos de selección positiva y negativa, tanto en linfocitos Tαβ, Tγδ, B1 y B2. • Describir la circulación y distribución de estas poblaciones celulares, a través de los sistemas circulatorio y linfático.	1 – 3	6	4	2	-
2	Respuesta inmune innata	Barreras físicas, químicas y biológicas	• Definir el concepto de inmunidad natural. • Describir los distintos componentes que integran los mecanismos naturales de defensa en el organismo. • Reconocer y describir la importancia de las barreras físicas, químicas y microbiológicas.	1 – 2	3	3	-	-

Universidad Nacional Autónoma de México
Facultad de Medicina
Departamento de
Tabla de Especificaciones de Inmunología – Año escolar 2014

Facultad de Medicina

Unid	Contenido (Tema)	Subtema	Objetivo temático	Ponderación (sesiones-horas)	Subtotal de reactivos	Nivel de desempeño		
						Conocimiento	Comprensión	Aplicación
		Reconocimiento en la respuesta inmune innata	• Describir los patrones moleculares asociados a patógenos (PAMPs) y asociados a daño (DAMPs). • Describir los receptores de reconocimiento de patrones (PRRs): tipo Toll (TLRs), tipo NOD (NLRs), tipo RIG (RLRs) lectinas solubles y transmembranales (CLRs), Scavenger, pentraxinas, receptores de N-formilmetionina, receptor para LPS, receptores para opsoninas.	1 – 3	6	6	-	-
		Respuesta inflamatoria	• Describir las citocinas y quimiocinas que participan en la respuesta inflamatoria. • Describir las características moleculares y de función de las familias de moléculas de adhesión que participan en la respuesta inflamatoria. • Describir el papel de los endotelios activados. • Describir a las células inmunes que participan en la respuesta inflamatoria.	1 – 3	6	4	2	-
		Fagocitosis	• Reconocer y describir a la fagocitosis como un mecanismo temprano e inespecífico de protección. • Mencionar los efectos quimiotácticos, de adhesión, de endocitosis, la formación del fagosoma y el lisosoma y su fusión. • Describir las poblaciones de células fagocíticas. • Describir la activación de las células fagocíticas a través de receptores de superficie. • Describir los mecanismos oxígeno-dependientes, oxígeno-independientes y la formación de óxido nítrico que intervienen en la eliminación de microorganismos. • Reconocer las consecuencias de la fagocitosis (destrucción del agente extraño y procesamiento de antígeno).	1 – 2	4	2	2	-
		Sistema del complemento	• Describir la naturaleza de los componentes del sistema del complemento • Describir la secuencia de activación de los componentes del sistema del complemento que tienen lugar, tanto en la vía clásica, la vía alterna y la vía de las lectinas. • Describir los mecanismos de daño ocasionados a la célula blanco por los complejos moleculares del sistema del complemento. • Reconocer sus mecanismos de regulación, tanto solubles como de membrana, y definir su importancia para una óptima respuesta inmune, así como en la prevención de daño a células y estructuras propias. • Describir los receptores para el complemento y su función. • Describir las funciones biológicas del complemento: lisis, opsonización, inflamación, activación celular, solubilización de complejos inmunes.	2 – 5	7	4	3	-

Universidad Nacional Autónoma de México
Facultad de Medicina
Departamento de
Tabla de Especificaciones de Inmunología – Año escolar 2014

Facultad de Medicina

Unid	Contenido (Tema)	Subtema	Objetivo temático	Ponderación (sesiones-horas)	Subtotal de reactivos	Nivel de desempeño		
						Conocimiento	Comprensión	Aplicación
3	Respuesta inmune adaptativa	Antígenos	• Reconocer las características estructurales que distinguen a los antígenos, inmunógenos y haptenos. • Describir el concepto de determinante antigénico y las características de los determinantes antigénicos secuenciales, conformacionales, ocultos e inmunodominantes. • Describir las características del antígeno necesarias para que exista antigenicidad, así como las características biológicas de los organismos necesarias para una respuesta inmune adecuada. • Reconocer las diferencias entre inmunógeno, mitógeno y superantígeno.	1 – 2	2	2	-	-
		Anticuerpos	• Definir la estructura de las inmunoglobulinas, señalando las diferentes clases y subclases y las cadenas pesadas y ligeras. • Definir la cadena J y el componente secretorio. • Señalar las diferencias entre isotipo, alotipo e idiotipo. • Describir las clases de receptores para el Fc de las inmunoglobulinas y su papel biológico. • Reconocer las teorías que se han postulado para explicar el origen de la diversidad, hasta el conocimiento actual gracias a las técnicas de biología molecular. • Describir los diferentes mecanismos moleculares (rearreglos de genes, diversidad de unión, hipermutación somática) que originan la diversidad de especificidades de anticuerpos y su ensamblaje. • Describir los cambios de clase de las inmunoglobulinas. • Señalar las funciones biológicas de las inmunoglobulinas. • Reconocer las metodologías empleadas para la obtención de anticuerpos monoclonales por hibridomas y por ingeniería genética. • Reconocer las principales aplicaciones de las inmunoglobulinas en el diagnóstico y en la terapéutica.	1 – 3	6	4	2	-
		Reacción antígeno-anticuerpo	• Describir qué es una reacción antígeno-anticuerpo, definiendo los conceptos de especificidad, afinidad, valencia y avidez de los anticuerpos. • Señalar los tipos de unión que participan y el sentido biológico de estos. • Describir las condiciones fisicoquímicas que afectan la unión antígeno-anticuerpo. • Explicar la reacción de precipitación antígeno-anticuerpo en términos moleculares. Describir la curva de precipitación. • Señalar las diferencias entre antígenos particulares y antígenos solubles y la diferencia entre precipitación y aglutinación. • Reconocer los principios en que se basan las distintas técnicas de aglutinación, precipitación y de inhibición, así como los distintos métodos (ELISA, radioinmunoensayo, inmunofluorescencia, etc.) y sus aplicaciones en el diagnóstico.	1 – 2	2	1	1	-

Universidad Nacional Autónoma de México
Facultad de Medicina
Departamento de
Tabla de Especificaciones de Inmunología – Año escolar 2014

Facultad de Medicina

Unid	Contenido (Tema)	Subtema	Objetivo temático	Ponderación (sesiones-horas)	Subtotal de reactivos	Nivel de desempeño		
						Conocimiento	Comprensión	Aplicación
		Procesamiento y presentación de antígenos	• Definir el complejo principal de histocompatibilidad (MHC), las moléculas de clase I y II, clásicas y no clásicas, y de clase III. • Describir las moléculas codificadas por los genes de las regiones del MHC clase III. • Describir la estructura, función y polimorfismo del MHC. • Describir la organización genética del MHC. • Describir las principales características de las moléculas de la familia CD1. • Describir los mecanismos de procesamiento y presentación del antígeno, así como el tipo de antígeno presentado, ya sea endógeno o exógeno. • Describir el procesamiento de antígenos lipídicos y su presentación por CD1.	2 – 5	10	6	4	-
		Receptores para antígeno	• Describir las características moleculares de los receptores para antígeno en los linfocitos B y T (BCR y TCR), sus regiones génicas y los rearreglos que dan origen al repertorio en el caso del TCR. • Describir los complejos moleculares asociados a los receptores: CD79α, CD79β, CD19 y CD21 para el BCR, y CD3 y CD247 (cadenas ζ) para el TCR. • Describir el proceso de reconocimiento del antígeno por los linfocitos T y B. • Definir el concepto de cooperación celular • Describir las principales moléculas coestimuladoras para linfocitos B (CD80 y CD86, CD40, CD19, CD21 y CD81) • Describir las principales moléculas coestimuladoras para linfocitos T (CD2, CD4, CD8, CD28, CTLA-4 e ICOS). • Describir las moléculas de adhesión que participan en los procesos de cooperación celular.	1 – 2	4	4	-	-
		Activación de linfocitos T y B	• Definir el concepto de cooperación celular. • Describir las cascadas de activación de linfocitos T y B. • Describir la activación de linfocitos B en los focos extrafoliculares y en el folículo linfoide. • Describir los mecanismos de cambio de isotipo y de hipermutación somáica. • Describir las características de la generación de células plasmáticas y células de memoria.	2 – 5	8	6	1	1
		Mecanismos de citotoxicidad	• Describir los mecanismos de reconocimiento y efectores de las células NK y reconocer la importancia de estos eventos en la destrucción de células tumorales, bajo estrés o infectadas por virus. • Describir los mecanismos efectores de los linfocitos $TCD8^{+}$. • Describir las características y funciones efectoras de los linfocitos Tγδ.	1 – 2	3	2	1	-

Universidad Nacional Autónoma de México
Facultad de Medicina
Departamento de
Tabla de Especificaciones de Inmunología – Año escolar 2014

Facultad de Medicina

Unid	Contenido (Tema)	Subtema	Objetivo temático	Ponderación (sesiones-horas)	Subtotal de reactivos	Nivel de desempeño		
						Conocimiento	Comprensión	Aplicación
			• Describir las células, los isotipos de inmunoglobulinas y los mecanismos involucrados en la citotoxicidad celular dependiente de anticuerpos.					
4	Regulación de la respuesta inmune	Citocinas	• Identificar las características estructurales y funcionales de las diversas familias de receptores de citocinas. • Describir las características principales de las citocinas (redundancia, pleiotropismo, sinergia y antagonismo) • Reconocer y describir la importancia de las citocinas de acuerdo a su función: • Proinflamatorias (respuesta inmune innata): IL-1, IL-6, IL-15, IL-17, IL-18, TNF-α, IFN-α y β, IFN-γ. • Respuesta inmune adaptativa: IL-2, IL-4, IL-5, IL-10, IL-12, IL-13, IFN-γ, TGF-β. • Hematopoyéticas: IL-3, IL-7, IL-11, GM-CSF, G-CSF, M-CSF, c-Kit ligando, eritropoyetina. • Quimiocinas: IL-8, eotaxina, MCP-1, 2 y 3, RANTES, MIP-1, CCL-19, CCL-21, CXCL-13.	2 – 5	10	8	2	-
		Subpoblaciones de linfocitos T	• Describir el proceso de modulación de la respuesta inmune a través del paradigma Th1/Th2. • Describir las características, funciones y señales necesarias para la activación de las distintas subpoblaciones de linfocitos T (Th1, Th2, Treg, Th17). • Describir las características funcionales de las diversas poblaciones linfocitarias reguladoras (nTreg, Th3, Tr1, iTr35), y de otras subpoblaciones (Th5, Th9, Th22).	1 – 3	6	4	2	-
		Otros mecanismos de regulación	• Definir el concepto e importancia de la regulación de la respuesta inmune. • Describir los mecanismos de regulación por anticuerpos, por antígeno, por complejos inmunes, por idiotipo-antiidiotipo y por factores supresores. • Describir el proceso de apoptosis, los mecanismos que participan, así como las señales necesarias para su activación. • Describir las características funcionales de otras células con funciones reguladoras: macrófagos M2 (alternativos) y M3 (reguladores), linfocitos B reguladores. • Describir las interacciones entre el sistema inmune y el sistema neuroendocrino.	1 – 2	0	-	-	-

Universidad Nacional Autónoma de México
Facultad de Medicina
Departamento de
Tabla de Especificaciones de Inmunología – Año escolar 2014

Facultad de Medicina

Unid	Contenido (Tema)	Subtema	Objetivo temático	Ponderación (sesiones-horas)	Subtotal de reactivos	Nivel de desempeño		
						Conocimiento	Comprensión	Aplicación
5	Introducción a la Inmunología clínica e Inmunopato-logía	Respuesta inmune y agentes infecciosos	• Describir el papel de la respuesta inmune ante la invasión de patógenos. • Describir los mecanismos efectores de las barreras, de la inmunidad innata y de la adaptativa que se monta para la protección contra virus, bacterias (intracelulares y extracelulares), hongos y parásitos. • Identificar el tipo de respuesta predominante ante cada tipo de agente infeccioso (virus, bacterias, hongos y parásitos). • Describir los mecanismos de evasión de la respuesta inmune por parte de los distintos agentes patógenos.	2 – 5	10	6	4	-
		Bases biológicas de la vacunación	• Describir las principales características de una vacuna. • Describir la importancia de los adyuvantes. • Identificar los principales tipos de vacunas.	1 – 2	0	-	-	-
		Mecanismos de daño por el sistema inmunológico: Hipersensibilidad	• Identificar las reacciones de hipersensibilidad de acuerdo a la clasificación de Gell y Coombs. • Describir los mecanismos de la respuesta inmunitaria que ocurren en la producción de daño en las reacciones de hipersensibilidad, así como las distintas fases y clasificaciones dentro de cada una. • Relacionar los mecanismos de daño con las principales entidades patológicas en las que estos juegan un papel fundamental.	2 – 6	12	6	6	-
		Tolerancia y Autoinmunidad	• Reconocer el fenómeno de autoinmunidad fisiológica. • Describir los mecanismos de selección positiva y negativa en médula ósea y timo. • Identificar las alteraciones en la tolerancia debido a mutaciones en los genes AIRE y FOXP3 (síndromes poliglandulares autoinmunes). • Describir los mecanismos de generación de torelancia periférica, así como la autorreactividad. • Identificar los mecanismos que ocurren para romper la tolerancia. • Reconocer la importancia de los mecanismos de autoinmunidad y su asociación con antígenos HLA. • Definir el concepto de riesgo relativo, reconociendo las bases moleculares de dicha asociación. • Describir las principales enfermedades autoinmunes, tanto órgano específicas como sistémicas, ilustrando con ejemplos típicos, señalando las características más relevantes de la alteración de la respuesta inmune (enfermedad de Graves, tiroiditis de Hashimoto, miastenia gravis, anemia perniciosa, gastritis atrófica inmune, anemia hemolítica autoinmune, púrpura trombocitopénica, cirrosis biliar primaria, diabetes tipo I,	2 – 6	12	6	6	-

Universidad Nacional Autónoma de México
Facultad de Medicina
Departamento de
Tabla de Especificaciones de Inmunología – Año escolar 2014

Facultad de Medicina

Unid	Contenido (Tema)	Subtema	Objetivo temático	Ponderación (sesiones-horas)	Subtotal de reactivos	Nivel de desempeño		
						Conocimiento	Comprensión	Aplicación
			enfermedad de Addison, síndrome de Goodpasture, esclerosis múltiple, lupus eritematoso sistémico, artritis reumatoide, poliomiositis, dermatomiosistis, esclerodermia) • Relacionar los mecanismos de autoinmunidad con las reacciones de hipersensibilidad. • Identificar y describir las características de los sitios inmunológicamente privilegiados.					
		Inmunodeficiencias primarias	• Describir el proceso de maduración inmunológica en la niñez. • Definir el concepto de inmunodeficiencia. • Diferenciar las inmunodeficiencias primarias de las secundarias. • Clasificar los tipos de inmunodeficiencias en relación al tipo de respuesta involucrada. • Describir la etiología y las características más relevantes de las inmunodeficiencias primarias de linfocitos B (agammaglobulinemia ligada al X, deficiencia selectiva de IgA y de subclases de IgG, inmunodeficiencia común variable, síndrome de hiper IgM). • Describir la etiología y las características más relevantes de las inmunodeficiencias primarias de linfocitos T (síndrome de Di George, candidiasis mucocutánea crónica, síndrome del linfocito desnudo). • Describir la etiología y las características más relevantes de las inmunodeficiencias combinadas graves (ligada al X y formas autosómicas recesivas) • Describir la etiología y las características más relevantes de las inmunodeficiencias primarias del sistema del complemento (deficiencia de componentes de la cascada de activación, deficiencia de proteínas reguladoras solubles y de membrana, deficiencias en los receptores). • Describir la etiología y las características más relevantes de las inmunodeficiencias primarias de la fagocitosis (enfermedad granulomatosa crónica, deficiencia de glucosa-6-fosfato deshidrogenasa o mieloperoxidasa, síndrome de Chédiak-Higashi, deficiencias de adhesión leucocitaria).	2 – 5	10	6	4	-
		Inmunodeficiencias secundarias	• Definir el concepto de inmunodeficiencia secundaria. • Describir el mecanismo de infección por el VIH, los mecanismos que participan en el daño de las células inmunocompetentes, los antígenos virales y su importancia en el diagnóstico y evolución del padecimiento, los recursos terapéuticos empleados actualmente contra este virus. • Describir el papel de la desnutrición grave y crónica en la alteración de la respuesta inmune. • Identificar otras causas de inmunodeficiencia secundaria.	1 – 2	4	2	2	-

Universidad Nacional Autónoma de México
Facultad de Medicina
Departamento de
Tabla de Especificaciones de Inmunología – Año escolar 2014

Facultad de Medicina

Unid	Contenido (Tema)	Subtema	Objetivo temático	Ponderación (sesiones-horas)	Subtotal de reactivos	Nivel de desempeño		
						Conocimiento	Comprensión	Aplicación
		Respuesta inmune y cáncer	• Definir el concepto de neoplasia. • Describir el concepto de vigilancia inmunológica. • Identificar los distintos tipos de antígenos tumorales, sus características y su importancia en el diagnóstico y pronóstico de la enfermedad. • Describir los principales virus oncogénicos. • Describir los distintos tipos de respuesta inmune ante una neoplasia, y la infiltración del tumor por células inmunocompetentes. • Describir los mecanismos de evasión de la respuesta inmune desarrollados por el tumor.	1 – 3	6	3	2	1
		Respuesta inmune y reproducción	• Describir las características de la respuesta inmune en los tractos genitales femenino y masculino. • Identificar el papel de la placenta como órgano con actividad inmunitaria. • Describir el papel de citocinas y hormonas en la tolerancia inmunológica al embrión y al feto. • Describir el papel del MHC en la inducción de tolerancia. • Describir los mecanismos de protección materno-fetal.	1 – 2	0	-	-	-
		Respuesta inmune a trasplantes	• Describir las características y los tipos de trasplante. • Describir el concepto de compatibilidad. • Describir los distintos mecanismos inmunológicos involucrados en el rechazo de los órganos o tejidos transplantados, y la enfermedad injerto contra huésped. • Identificar las diferentes medidas terapéuticas utilizadas para prevenir el rechazo del injerto.	1 – 2	0	-	-	-
Totales				34 - 84	140	92	46	2

Anexo 3. Comparación de recomendaciones para elaboración de reactivos

Tipo	Haladyna, Downing, Rodríguez (traducción del Manual para elaboración de pruebas a gran escala, CENEVAL)	Haladyna, Downing, Rodríguez (traducción de Moreno, Martínez y Muñiz, 2004)	Swanson, Case (Cómo construir preguntas de Selección Múltiple para Ciencias Básicas y Ciencias Clínicas, 2005)	Moreno, Martínez, Muñiz (Directrices para la construcción de ítems de opción múltiple, 2004)	Amin, Z (Practical Guide to Medical Student Assessment, 2006)	Otras propuestas.
Contenido	Todo reactivo debe de reflejar un contenido específico y una sola conducta mental, como sea requerido en las especificaciones de la prueba	Cada ítem debe de reflejar un contenido específico y una única conducta mental específica, tal como sea requerido en las especificaciones del test (tabla de doble entrada, proyecto del test)		La representatividad deberá marcar lo sencillo o complejo, concreto o abstracto, memorístico o de razonamiento que deba ser el ítem, así como el modo de expresarlo	Se enfoca en un aspecto (indicación, efectos adversos, contraindicación, mecanismo de acción).	Cada ítem debe enfocarse en un único problema o idea (Zimmaro) Solamente debe evaluar un resultado de aprendizaje. (INB) Incluir una sola idea al elaborar el reactivo, es decir, presentar solamente un problema (CENEVAL) Alinear los reactivos con la tabla de contenidos o con las especificaciones de reactivos (CENEVAL)
	Basar cada reactivo en contenido importante, evitar contenidos triviales	Base cada ítem en un contenido importante para el aprendizaje; evite contenidos triviales	Cada ítem debiera enfocar un concepto importante, típicamente un problema clínico común o potencialmente catastrófico.	Debe ser una muestra representativa del contenido recogido en una tabla de especificación, evitando ítems triviales	El tópico es importante para los estudiantes.	No examinar contenidos intrascendentes o triviales (CENEVAL)
	Usar material novedoso para evaluar aprendizaje del nivel más alto. No usar el mismo lenguaje que en el libro de texto, para evitar evaluar el recuerdo de palabras y oraciones familiares	Use material novedoso para evaluar el aprendizaje de alto nivel. Cuando los utilice en un ítem, parafrasee el lenguaje de los libros de texto, o el lenguaje utilizado durante la instrucción, para así evitar evaluar el mero recuerdo.	Cada ítem debiera evaluar la aplicación del conocimiento y no el recuerdo de un hecho aislado		Evalúa más allá del recuerdo del conocimiento y la memorización.	Evitar conceptos citados de manera textual. (CENEVAL)
	Mantener independiente el contenido de cada reactivo	Mantenga el contenido de cada ítem independiente del contenido de otros ítems del test			El nivel de dificultad es adecuado.	
	Evitar contenido demasiado específico o demasiado general	Al escribir ítems de elección múltiple, evite contenidos muy específicos o muy generales				
	Evitar reactivos basados en opiniones	Evite ítems basados en opiniones				
	Evitar reactivos engañosos	Evite ítems con trampas				
	Mantener el vocabulario simple y adecuado a los individuos que serán evaluados	Use un vocabulario sencillo para el grupo de estudiantes que están siendo evaluados		La semántica debe estar ajustada al contenido y a las personas evaluadas	No contiene jerga o abreviaturas.	

Tipo	Haladyna, Downing, Rodríguez (traducción del Manual para elaboración de pruebas a gran escala, CENEVAL)	Haladyna, Downing, Rodríguez (traducción de Moreno, Martínez y Muñiz, 2004)	Swanson, Case (Cómo construir preguntas de Selección Múltiple para Ciencias Básicas y Ciencias Clínicas, 2005)	Moreno, Martínez, Muñiz (Directrices para la construcción de ítems de opción múltiple, 2004)	Tarrant, M (Nurse Education Today, 2006)	Amin, Z (Practical Guide to Medical Student Assessment, 2006)	Otras propuestas.
Formato	Usar las versiones de pregunta, respuesta y llenado de los reactivos convencionales de opción múltiple, opción alterna, verdadero-falso, múltiples verdadero-falso, igualación, reactivo y grupo de reactivos dependientes de contexto; pero evitar el formato complejo de opción múltiple	Del formato convencional de elección múltiple utilice la interrogación, completar frases, la mejor respuesta, elección alternativa, verdadero-falso, verdadero-falso múltiple, emparejamiento, los conjuntos de ítems y los dependientes de contexto; sin embargo, evite el formato de elección múltiple complejo (el tipo K)			6. Evite ROM complejos, o tipo K. ROM tipo K tienen un rango de respuestas correctas y se solicita a los estudiantes que elijan de entre un número de combinaciones posibles de estas respuestas. Los estudiantes pueden frecuentemente adivinar al eliminar una respuesta incorrecta de todas las opciones que incluyan esta respuesta o al seleccionar las respuestas que aparezcan más frecuentemente en todas las opciones.	Ítems basados en contexto tienen integrada la viñeta clínica.	
	Formatear los reactivos de manera vertical, no horizontal	Construya el ítem de forma vertical, no horizontal		Las opciones deben presentarse usualmente en vertical			
Estilo	Revisar y editar los reactivos	Corrija y pruebe los ítems					
	Usar gramática, puntuación, ortografía y mayúsculas de manera correcta	Utilice una gramática, puntuación, mayúsculas y minúsculas y deletreo correctos		La sintaxis o estructura gramatical debe ser correcta. Evitar ítems demasiado escuetos o profusos, ambiguos o confusos, cuidando además las expresiones negativas			Utilice una gramática, puntuación y sintaxis correctas de manera consistente. (Zimmaro)
	Minimizar la cantidad de lectura requerida para cada reactivo	Minimice la cantidad de lectura en cada ítem			5. Evite información arbitraria o innecesaria in el tallo o las opciones. Si se incluye una viñeta en el ROM, debe ser necesaria para responder la pregunta.		

Tipo	Haladyna, Downing, Rodríguez (traducción del Manual para elaboración de pruebas a gran escala, CENEVAL)	Haladyna, Downing, Rodríguez (traducción de Moreno, Martínez y Muñiz, 2004)	Swanson, Case (Cómo construir preguntas de Selección Múltiple para Ciencias Básicas y Ciencias Clínicas, 2005)	Moreno, Martínez, Muñiz (Directrices para la construcción de ítems de opción múltiple, 2004)	Tarrant, M (Nurse Education Today, 2006)	Amin, Z (Practical Guide to Medical Student Assessment, 2006)
Redacción del tallo	Asegurarse de que las instrucciones en la base sean muy claras	Asegúrese de que el sentido del enunciado resulta muy claro			7. Las preguntas y todas las opciones deben ser escritas en un lenguaje claro, no ambiguo. Un lenguaje pobre o preguntas ambiguas pueden confundir incluso a los estudiantes conocedores y causarles que respondan de manera incorrecta	El tallo es claro y completo.
	Incluir la idea central en la base, no en las opciones	Incluya la idea central en el enunciado y no en las opciones	El enunciado de los ítems debe incluir una pregunta hecha en forma clara y debe ser posible dar la respuesta correcta sin mostrar las opciones.	Lo central debe expresarse en el enunciado. Cada opción es un complemento que debe concordar gramaticalmente con el enunciado	2. Cada ROM debe tener una pregunta clara y enfocada. Los profesores deben evitar usar ROM con tallos no enfocados los cuales no hagan una pregunta clara o enuncien un problema claro en el formato de completamiento. 3. Cada ROM debe tener el problema en el tallo de la pregunta, no en las opciones. Las opciones no deben de ser una serie de enunciados verdadero/falso.	Se puede contestar sin mirar las opciones.
	Evitar adornar el reactivo (evite el exceso de palabras)	Evite adornar el texto en exceso (palabrería excesiva)				
	Plantear la base de manera positiva, evitar negativos como NO o EXCEPTO. Si se usan palabras negativas, usarlas de manera cuidadosa y siempre asegurarse que la palabra se inicia con mayúsculas y negritas	Exprese el enunciado de manera afirmativa, evitando términos negativos tales como NO o EXCEPTO. Si usa términos negativos, hágalo con sumo cuidado y asegúrese que aparecen en mayúsculas o negritas			16. Evite el uso de negativos (no, excepto, incorrecto) en el tallo ya que evalúa mal el conocimiento real de los estudiantes. Si los profesores desean evaluar contraindicaciones, la pregunta debe ser redactada claramente para indicar que esto es lo que será evaluado.	

Tipo	Haladyna, Downing, Rodríguez (traducción del Manual para elaboración de pruebas a gran escala, CENEVAL)	Haladyna, Downing, Rodríguez (traducción de Moreno, Martínez y Muñiz, 2004)	Swanson, Case (Cómo construir preguntas de Selección Múltiple para Ciencias Básicas y Ciencias Clínicas, 2005)	Moreno, Martínez, Muñiz (Directrices para la construcción de ítems de opción múltiple, 2004)	Tarrant, M (Nurse Education Today, 2006)	Amin, Z (Practical Guide to Medical Student Assessment, 2006)
Redacción de las opciones (1)	Desarrollar tantas opciones efectivas como sea posible, pero recordar que los resultados de investigación sugiere que tres son adecuadas	Escriba tantas opciones como pueda, si bien la investigación sugiere que con tres es suficiente		Las opciones deben ser preferiblemente tres		
	Asegurarse de que únicamente una de esas opciones es la correcta	Asegúrese que sólo una de esas opciones es la respuesta correcta		La opción correcta debe ser sólo una, acompañada por distractoras plausibles	4. El formato básico de los ROM es el de mejor respuesta única. Por ello, asegúrese de que las preguntas tengan una, y sólo una, mejor respuesta.	
	Variar la posición de la respuesta correcta de acuerdo al número de opciones. Equilibrar la clave de respuestas dentro de lo posible, de modo que la respuesta correcta aparezca el mismo número de veces en cada posición	Varíe la colocación de la respuesta correcta según el número de opciones		La opción correcta debe estar repartida entre las distintas ubicaciones		
	Colocar las opciones en orden lógico o numérico	Coloque las opciones en un orden lógico o numérico		El conjunto de opciones de cada ítem debe aparecer estructurado	13. Organice las opciones de los ROM en orden alfabético, cronológico o numérico.	
	Mantener las opciones independientes; las opciones no deben superponerse en significado	Construya las opciones independientes entre sí, no deben solaparse		Las opciones deben ser autónomas entre sí, sin solaparse ni referirse unas a otras. Por ello, deben evitarse las opciones «Todas las anteriores» y «Ninguna de las anteriores»		
	Mantener a las opciones homogéneas en estructura gramatical y contenido	Mantenga a las opciones homogéneas en contenido y estructura gramatical	Los distractores (las opciones incorrectas) debieran ser homogéneos	Ninguna opción debe destacar del resto ni en contenido ni en apariencia	10. Evite ofrecer pistas lógicas en el tallo y la opción correcta que puedan ayudar al estudiante a identificar la opción correcta sin conocer el material. Un ejemplo de una pista lógica es solicitar a los estudiantes que seleccionen la intervención farmacológica más adecuada para un problema y solamente tener una o dos opciones la cuales sean realmente intervenciones farmacológicas.	Todas las opciones son uniformes (tamaño, construcción gramatical)
	Igualar la extensión de todas las opciones	Escriba las opciones con una longitud aproximadamente igual			12. Todas las opciones deben ser similares en extensión y cantidad de detalle provisto en la opción. Si una opción es más grande, incluye información más detallada o contiene un lenguaje más complejo, los estudiantes usualmente asumen correctamente que esa es la respuesta correcta.	
	Ninguna de las anteriores debe usarse con cuidado	La opción *Ninguno de las anteriores* debe usarse con mucho cuidado			18. Evite el uso de "ninguna de las anteriores" como la última opción ya que sólo mide la habilidad de los estudiantes de detectar respuestas incorrectas. Además, si "ninguna de las anteriores" es la opción correcta, el profesor debe tener la certeza de que no hay excepciones a ninguna de las opciones que el estudiante haya detectado.	No hay opciones como "todas las anteriores" o "ninguna de las anteriores".
	Evitar el uso de *Todas las anteriores*	Evite la opción *Todas las anteriores*			17. Evite el uso de "todas las anteriores" como la última opción. Los estudiantes pueden fácilmente identificar si esta es la opción correcta únicamente sabiendo que al menos dos de las opciones son correctas. De igual modo, pueden anularla si saben que sólo una de las opciones es incorrecta.	
	Plantear las opciones de forma positiva; evitar negativos como no	Escriba las opciones de forma afirmativa, evite términos negativos tales como NO				

Tipo	Haladyna, Downing, Rodríguez (traducción del Manual para elaboración de pruebas a gran escala, CENEVAL)	Haladyna, Downing, Rodríguez (traducción de Moreno, Martínez y Muñiz, 2004)	Swanson, Case (Cómo construir preguntas de Selección Múltiple para Ciencias Básicas y Ciencias Clínicas, 2005)		Tarrant, M (Nurse Education Today, 2006)	Amin, Z (Practical Guide to Medical Student Assessment, 2006)	Otras propuestas
Redacción de las opciones (2)	Evitar dar claves o pistas de la respuesta correcta, tales como: a. Determinantes específicos como siempre, nunca, completamente y absolutamente b. Asociaciones "que suenan", opciones idénticas o parecidas a la base c. Inconsistencias gramaticales que dan pistas de la respuesta correcta d. Opciones flagrantemente correctas e. Pares o tríos de opciones que dan pistas de la respuesta correcta f. Opciones flagrantemente absurdas y ridículas	Evite dar pistas sobre la respuesta correcta, tales como: a. Determinantes específicos como siempre, nunca, completamente y absolutamente b. Asociaciones por sonido similar y opciones idénticas o parecidas a términos del enunciado c. Inconsistencias gramaticales que indiquen al sujeto la elección correcta d. Opción correcta destacada e. Pares o tríos de opciones que indiquen al sujeto la opción correcta f. Opciones claramente absurdas o ridículas	Evite defectos de los ítems que beneficien a aquellos examinados que son astutos en percibir los defectos de construcción y así guiarse hacia la respuesta correcta. Evite asimismo defectos que resultan en preguntas difíciles sólo por fallas en su construcción.		15. Las opciones deben redactarse evitando el uso de términos absolutos (nunca, siempre, únicamente, todos) ya que a los estudiantes se les enseña que usualmente no hay verdades absolutas en la mayoría de los temas de ciencias de la salud y pueden entonces eliminar estos distractores. 1. Todas las opciones deben ser gramáticamente consistentes con el tallo y deben ser similares en estilo y forma. Opciones con inconsistencias gramaticales proveen pistas a los estudiantes, quienes fácilmente eliminan distractores que no concuerdan gramaticalmente con el tallo. 9. Evite repetir palabras en el tallo y en la opción correcta. Palabras similares permiten al estudiante reconocer la opción correcta sin conocer el material. 11. Evite pistas convergentes en las opciones donde hay diferentes combinaciones de múltiples componentes de la respuesta. Los elaboradores de reactivos tienden a usar las respuestas correctas más frecuentemente entre las opciones y los estudiantes identifican como correcta la respuesta en la cual todos los componentes aparecen más frecuentemente.	Las opciones no dan pista de la respuesta No se utilizan términos ambiguos (casi, nunca, frecuente).	
	Hacer que todos los distractores sean verosímiles	Haga plausibles todos los distractores			8. Haga todos los distractores plausibles ya que distractores plausibles son vitales para ROM de alta calidad. Estudiantes que no saben el material aumentan sus probabilidades de adivinar la opción correcta al eliminar distractores no plausibles.		Deben ser plausibles, es decir, que no se descarten por inferencia lógica o sentido común. (CENEVAL)
	Usar errores típicos de los estudiantes para crear distractores	Use errores usuales de los estudiantes para escribir los distractores					Incluir los errores más comunes de los sustentantes. (CENEVAL)
	Usar humor si es compatible con el maestro y el ambiente de aprendizaje	Use el humor si es compatible con el profesor y con el ambiente de aprendizaje					
					15. Las opciones deben redactarse evitando el uso de términos imprecisos (frecuentemente, ocasionalmente, rara vez, usualmente, comúnmente) ya que estos términos carecen de precisión y pocas veces hay un acuerdo en el significado correcto de "frecuentemente".		
					19. Evite el formato de "complete el espacio" en donde una palabra se omite en medio de la oración y el estudiante debe adivinar la respuesta correcta. Todas las opciones deben ser colocadas al final del tallo.		

Anexo 4. Primera propuesta de instrumento para evaluar reactivos de opción múltiple con 24 indicadores

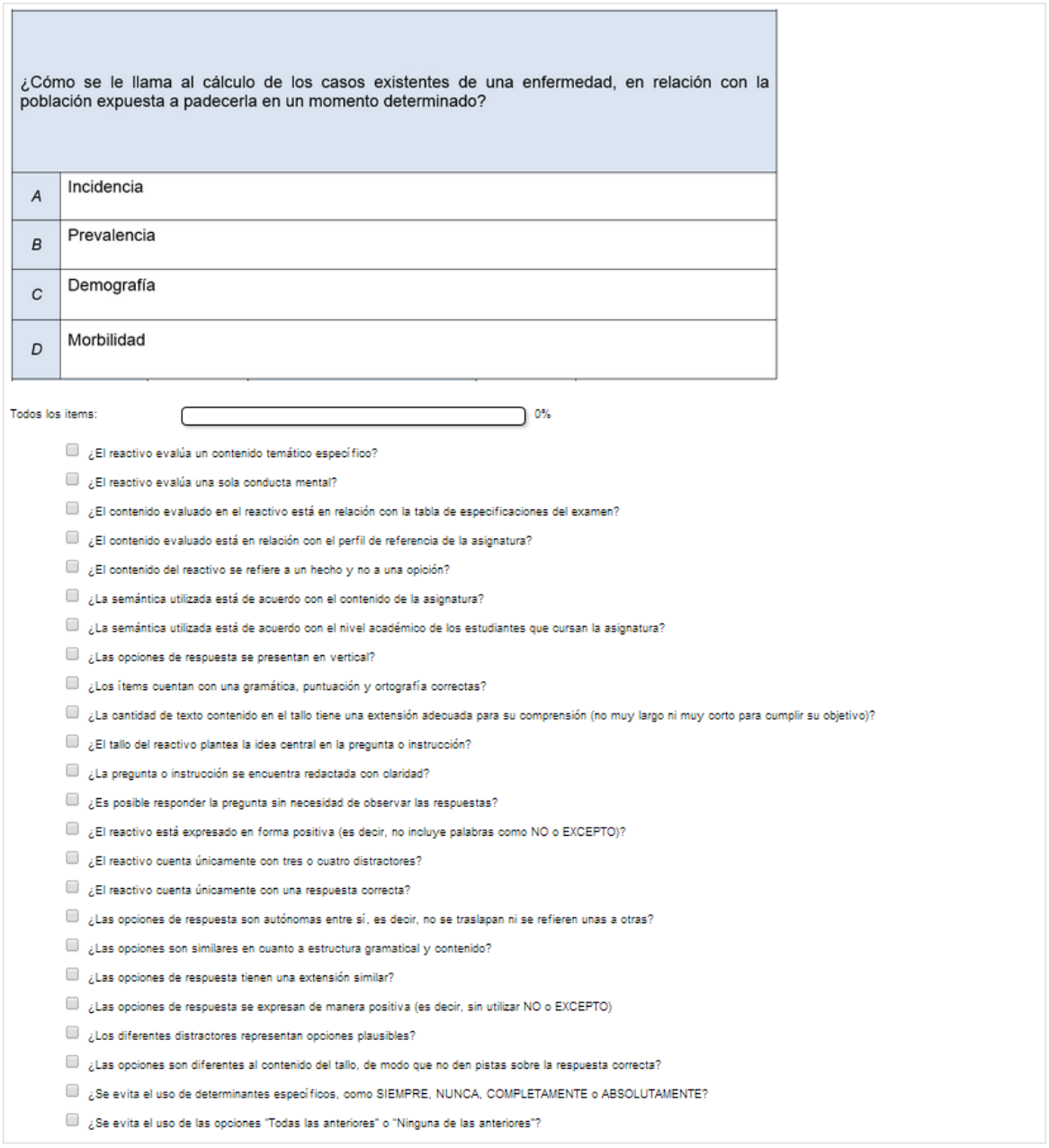
Instrumento para evaluar reactivos de opción múltiple

Este instrumento tiene como objetivo la evaluación cualitativa de los reactivos de opción múltiple (ROM) utilizados en los exámenes sumativos de la asignatura de Inmunología. Cada pregunta corresponde a una característica que se pretende encontrar en los ROM. Marque con una X si el ROM presentado cumple (Sí) o no cumple (No) con lo que se pregunta en cada uno de los enunciados.

Nivel taxonómico que evalúa: ( ) Conocimiento ( ) Comprensión ( ) Aplicación

		Si	No
	Contenido		
1	¿El reactivo evalúa un contenido temático específico?		
2	¿El reactivo evalúa una sola conducta mental?		
3	¿El contenido evaluado en el reactivo está en relación con la tabla de especificaciones del examen?		
4	¿El contenido evaluado en el reactivo está en relación con el perfil de referencia de la asignatura?		
5	¿El contenido del reactivo se refiere a un hecho y no a una opinión?		
6	¿La semántica utilizada está de acuerdo con el contenido de la asignatura?		
7	¿La semántica utilizada está de acuerdo con el nivel académico de los estudiantes que cursan la asignatura?		
	Formato y estilo		
8	¿Las opciones de respuesta se presentan en vertical?		
9	¿Los ítems cuentan con una gramática, puntuación y ortografía correctas?		
	Redacción del tallo		
10	¿La cantidad de texto contenido en el tallo tiene una extensión adecuada para su comprensión (no muy largo ni muy corto para cumplir su objetivo)?		
11	¿El tallo del reactivo plantea la idea central en la pregunta o instrucción?		
12	¿La pregunta o instrucción se encuentra redactada con claridad?		
13	¿Es posible responder la pregunta sin necesidad de observar las respuestas?		
14	¿El reactivo está expresado en forma positiva (es decir, no incluye palabras como NO o EXCEPTO)?		
	Redacción de las opciones		
15	¿El reactivo cuenta únicamente con tres o cuatro distractores?		
16	¿El reactivo cuenta únicamente con una respuesta correcta?		
17	¿Las opciones de respuesta son autónomas entre sí, es decir, no se traslapan ni se refieren unas a otras?		
18	¿Las opciones son similares en cuanto a estructura gramatical y contenido?		
19	¿Las opciones de respuesta tienen una extensión similar?		
20	¿Las opciones de respuesta se expresan de manera positiva (es decir, sin utilizar NO o EXCEPTO)		
21	¿Los diferentes distractores representan opciones plausibles?		
22	¿Las opciones son diferentes al contenido del tallo, de modo que no den pistas sobre la respuesta correcta?		
23	¿Se evita el uso de determinantes específicos, como SIEMPRE, NUNCA, COMPLETAMENTE o ABSOLUTAMENTE?		
24	¿Se evita el uso de las opciones “Todas las anteriores” o “Ninguna de las anteriores”?		

Anexo 5. Interfaz del cuestionario para obtener evidencia de validez por el juicio de expertos para el diseño del instrumento



¿Cómo se le llama al cálculo de los casos existentes de una enfermedad, en relación con la población expuesta a padecerla en un momento determinado?

A	Incidencia
B	Prevalencia
C	Demografía
D	Morbilidad

Todos los items: 0%

- ¿El reactivo evalúa un contenido temático específico?
- ¿El reactivo evalúa una sola conducta mental?
- ¿El contenido evaluado en el reactivo está en relación con la tabla de especificaciones del examen?
- ¿El contenido evaluado está en relación con el perfil de referencia de la asignatura?
- ¿El contenido del reactivo se refiere a un hecho y no a una opición?
- ¿La semántica utilizada está de acuerdo con el contenido de la asignatura?
- ¿La semántica utilizada está de acuerdo con el nivel académico de los estudiantes que cursan la asignatura?
- ¿Las opciones de respuesta se presentan en vertical?
- ¿Los ítems cuentan con una gramática, puntuación y ortografía correctas?
- ¿La cantidad de texto contenido en el tallo tiene una extensión adecuada para su comprensión (no muy largo ni muy corto para cumplir su objetivo)?
- ¿El tallo del reactivo plantea la idea central en la pregunta o instrucción?
- ¿La pregunta o instrucción se encuentra redactada con claridad?
- ¿Es posible responder la pregunta sin necesidad de observar las respuestas?
- ¿El reactivo está expresado en forma positiva (es decir, no incluye palabras como NO o EXCEPTO)?
- ¿El reactivo cuenta únicamente con tres o cuatro distractores?
- ¿El reactivo cuenta únicamente con una respuesta correcta?
- ¿Las opciones de respuesta son autónomas entre sí, es decir, no se traslapan ni se refieren unas a otras?
- ¿Las opciones son similares en cuanto a estructura gramatical y contenido?
- ¿Las opciones de respuesta tienen una extensión similar?
- ¿Las opciones de respuesta se expresan de manera positiva (es decir, sin utilizar NO o EXCEPTO)
- ¿Los diferentes distractores representan opciones plausibles?
- ¿Las opciones son diferentes al contenido del tallo, de modo que no den pistas sobre la respuesta correcta?
- ¿Se evita el uso de determinantes específicos, como SIEMPRE, NUNCA, COMPLETAMENTE o ABSOLUTAMENTE?
- ¿Se evita el uso de las opciones "Todas las anteriores" o "Ninguna de las anteriores"?

Anexo 6. Instrumento preliminar para la evaluación de reactivos de opción múltiple

Instrumento para evaluar reactivos de opción múltiple

Este instrumento tiene como objetivo la evaluación cualitativa de los reactivos de opción múltiple (ROM) utilizados en los exámenes sumativos de la asignatura de Inmunología. Cada pregunta corresponde a una característica que se pretende encontrar en los ROM. Marque con una X si el ROM presentado cumple (Sí) o no cumple (No) con lo que se pregunta en cada uno de los enunciados.

Nivel taxonómico que evalúa: ( ) Conocimiento ( ) Comprensión ( ) Aplicación

		Si	No
Contenido			
1	¿El reactivo presenta un solo contenido temático?		
2	¿El reactivo presenta un solo resultado de aprendizaje?		
3	¿El contenido evaluado está en relación con la especificación del reactivo?		
4	¿El contenido del reactivo se refiere a una evidencia y no a una opinión?		
5	¿La semántica utilizada está de acuerdo con el contenido del programa académico?		
Formato y estilo			
6	¿Las opciones de respuesta se presentan en vertical?		
7	¿El reactivo cuenta con una gramática, puntuación y ortografía correctas?		
Redacción del tallo			
8	¿La cantidad de texto en el tallo es adecuada para su comprensión?		
9	¿El tallo del reactivo plantea la idea central?		
10	¿La pregunta o instrucción se encuentra redactada con claridad?		
11	¿Es posible responder la pregunta sin necesidad de observar las respuestas?		
12	¿El reactivo está expresado en forma positiva (es decir, no incluye palabras como NO o EXCEPTO)?		
Redacción de las opciones de respuesta			
13	¿El reactivo cuenta con tres o cuatro opciones?		
14	¿El reactivo cuenta únicamente con una respuesta correcta?		
15	¿Las opciones son independientes entre sí?		
16	¿Las opciones son similares en cuanto a estructura gramatical, contenido y extensión?		
17	¿Las opciones se expresan de manera afirmativa?		
18	¿Los distractores son plausibles, es decir, no se descartan por inferencia lógica o sentido común?		
19	¿Las opciones evitan dar pistas sobre la respuesta correcta?		
20	¿Se evita el uso de términos como SIEMPRE, NUNCA, COMPLETAMENTE o ABSOLUTAMENTE?		
21	¿Se evita el uso de las opciones “Todas las anteriores” o “Ninguna de las anteriores”?		

Anexo 7. Programa académico del Curso-Taller "Elaboración de reactivos para la evaluación del aprendizaje de Inmunología"

UNIVERSIDAD NACIONAL AUTÓNOMA DE MÉXICO
FACULTAD DE MEDICINA

DEPARTAMENTO DE BIOQUÍMICA
COORDINACIÓN DE EVALUACIÓN DE INMUNOLOGÍA

CURSO - TALLER
ELABORACIÓN DE REACTIVOS PARA LA EVALUACION
DEL APRENDIZAJE DE INMUNOLOGÍA

Profesor titular:
Jesús Rivera Jiménez

Profesoras invitadas:
Marlette Lobato Valverde
Florina Gatica Lara

4. Contenidos

I. Conceptos esenciales en evaluación

a. Evaluación educativa
b. Tipos de evaluación

II. Evaluación en la Facultad de Medicina

a. Evaluación en la asignatura de Inmunología

III. Elaboración del examen departamental

a. Perfil de referencia
b. Tabla de especificaciones
c. Banco de reactivos

IV. Características y elaboración de diferentes tipos de reactivos:

a. Reactivos de opción múltiple (ROM)
   i. Definición y características
   ii. Recomendaciones para la elaboración de reactivos de opción múltiple
   iii. ROM convencionales
   iv. Respuesta alternativa
   v. Verdadero-falso (convencional y múltiple)
   vi. Emparejamiento
   vii. ROM complejos (tipo K)
   viii. Ítems dependientes de contexto
b. Reactivos de completamiento, respuesta corta y tipo ensayo

V. Validación de los resultados del examen departamental

a. Análisis de ítems
b. Efectos negativos del uso de ítems defectuosos

5. Metodología

El curso-taller se imparte en una modalidad *blended learning*.

Las sesiones presenciales se llevarán a cabo en las aulas del Departamento de Informática Biomédica de la Facultad de Medicina. Se trabajará con presentación de los temas por los profesores, así como trabajo individual y en pequeños grupos.

Para las sesiones en línea y la entrega de actividades se utilizará un entorno virtual de aprendizaje, en este caso será la plataforma Moodle en el servidor de las Coordinaciones de Inmunología del Departamento de Bioquímica, cuya dirección web es: https://bioq9c2.fmedic.unam.mx/moodle/

Este sitio funcionará de lunes a domingo, en un horario de 7:00 am a 2:00 am. A cada profesor se le entregará al inicio del curso un nombre de usuario y una contraseña para poder acceder al mismo. Las actividades programadas para el aula virtual se realizarán de manera individual. Cada semana se publicarán los temas y actividades a desarrollar mediante una guía didáctica. Es importante atender la programación para dar cumplimiento a los ejercicios

planeados para cada sesión. Se contará con un foro de trabajo para atender dudas e inquietudes sobre las actividades desarrolladas, así como un foro de socialización para la interacción entre los participantes.

6. Fechas y horarios

Semana	Fecha	Aula y horario
Semana 1	Jueves 20 de marzo	Sesión Presencial - Aula 2 de IB, de 16:00 a 19:00 Trabajo en línea - Aula virtual, horario libre (2 horas)
Semana 2	Jueves 27 de marzo	Trabajo en línea - Aula virtual, horario libre (5 horas)
Semana 3	Jueves 3 de abril	Trabajo en línea - Aula virtual, horario libre (5 horas)
Semana 4	Jueves 10 de abril	Sesión Presencial - Aula 2 de IB, de 16:00 a 19:00 Trabajo en línea - Aula virtual, horario libre (2 horas)

7. Evaluación

Para la acreditación del taller será necesario cumplir con lo siguiente:

Actividad	Valor
Asistencia y participación en las sesiones presenciales	10%
Cuatro reactivos de opción múltiple convencionales	40%
Cuatro reactivos de diferentes tipos	40%
Participación en los foros virtuales	10%

Anexo 8. Distribución de número de respuestas correctas contestando ciegamente para 60 ítems con 4 opciones de respuesta

Aciertos	Probabilidad	Probabilidad acumulada
0	0.00000003	0.00000003
1	0.00000064	0.00000067
2	0.00000627	0.00000694
3	0.00004042	0.00004736
4	0.00019199	0.00023935
5	0.00071677	0.00095613
6	0.00219014	0.00314627
7	0.0056318	0.00877807
8	0.01243689	0.02121496
9	0.02395252	0.04516748
10	0.04071929	0.08588677
11	0.06169589	0.14758267
12	0.08397497	0.23155763
13	0.10335381	0.33491144
14	0.11565783	0.45056927
15	0.118228	0.56879727
16	0.11083875	0.67963603
17	0.09562559	0.77526162
18	0.0761463	0.85140793
19	0.0561078	0.90751573
20	0.03834033	0.94585606
21	0.02434307	0.97019913
22	0.01438454	0.98458367
23	0.00792192	0.99250559
24	0.00407099	0.99657658
25	0.00195407	0.99853065
26	0.00087683	0.99940748
27	0.00036805	0.99977553
28	0.00014459	0.99992012
29	0.00005318	0.99997331
30	0.00001832	0.99999162

Aciertos	Probabilidad	Probabilidad acumulada
31	0.00000591	0.99999753
32	0.00000179	0.99999932
33	0.0000005	0.99999982
34	0.00000013	0.99999996
35	0.00000003	0.99999999
36	0.00000001	1
37	0	1
38	0	1
39	0	1
40	0	1
41	0	1
42	0	1
43	0	1
44	0	1
45	0	1
46	0	1
47	0	1
48	0	1
49	0	1
50	0	1
51	0	1
52	0	1
53	0	1
54	0	1
55	0	1
56	0	1
57	0	1
58	0	1
59	0	1
60	0	1

Anexo 9. Distribución de número de respuestas correctas contestando ciegamente para 70 ítems con 4 opciones de respuesta

Aciertos	Probabilidad	Probabilidad acumulada
0	0	0
1	0.00000004	0.00000004
2	0.00000048	0.00000053
3	0.00000364	0.00000417
4	0.00002033	0.0000245
5	0.00008945	0.00011395
6	0.00032301	0.00043696
7	0.00098441	0.00142137
8	0.00258408	0.00400545
9	0.00593382	0.00993927
10	0.01206544	0.02200471
11	0.02193716	0.04394187
12	0.03595256	0.07989443
13	0.05346792	0.13336235
14	0.0725636	0.20592595
15	0.09030137	0.29622731
16	0.10347032	0.39969763
17	0.10955681	0.50925444
18	0.10752798	0.61678242
19	0.0980957	0.71487811
20	0.08338134	0.79825946
21	0.06617567	0.86443513
22	0.04913042	0.91356555
23	0.03417768	0.94774323
24	0.02231043	0.97005367
25	0.01368373	0.9837374
26	0.00789446	0.99163186
27	0.00428835	0.99592021
28	0.00219523	0.99811543
29	0.00105976	0.9991752
30	0.00048278	0.99965798
31	0.00020765	0.99986563
32	0.00008436	0.99994999
33	0.00003238	0.99998236
34	0.00001175	0.99999411
35	0.00000403	0.99999814
36	0.00000131	0.99999944
37	0.0000004	0.99999984
38	0.00000012	0.99999996
39	0.00000003	0.99999999
40	0.00000001	1
41	0	1
42	0	1
43	0	1
44	0	1
45	0	1
46	0	1
47	0	1
48	0	1
49	0	1
50	0	1
51	0	1
52	0	1
53	0	1
54	0	1
55	0	1
56	0	1
57	0	1
58	0	1
59	0	1
60	0	1
61	0	1
62	0	1
63	0	1
64	0	1
65	0	1
66	0	1
67	0	1
68	0	1
69	0	1
70	0	1

Anexo 10. Correlaciones entre las parejas de ítems del instrumento para evaluar reactivos de opción múltiple

Correlaciones

		Crit1	Crit2	Crit3	Crit4	Crit5	Crit6	Crit7	Crit8	Crit9	Crit10	Crit11	Crit12	Crit13	Crit14	Crit15	Crit16	Crit17	Crit18	Crit19	Crit20	Crit21
Crit1	Correlación de Pearson	1	.704**	.250**	.078*	.260**	-.029	-.020	.100**	.102**	.063	.197**	.040	.020	-.007	.038	-.026	-.016	.055	-.014	c	-.010
	Sig. (bilateral)		.000	.000	.018	.000	.373	.546	.002	.002	.057	.000	.220	.538	.843	.252	.429	.621	.098	.672	.	.755
	N	924	924	924	924	924	924	924	924	924	924	924	924	924	924	924	924	924	924	924	924	924
Crit2	Correlación de Pearson	.704**	1	.214**	.066*	.225**	-.033	.042	.095**	.182**	.053	.152**	.060	.006	-.015	.057	-.027	-.019	.050	-.026	c	-.012
	Sig. (bilateral)	.000		.000	.045	.000	.311	.203	.004	.000	.110	.000	.067	.867	.646	.084	.408	.573	.126	.421	.	.722
	N	924	924	924	924	924	924	924	924	924	924	924	924	924	924	924	924	924	924	924	924	924
Crit3	Correlación de Pearson	.250**	.214**	1	.253**	.258**	-.020	.027	.187**	.261**	.148**	.041	.136**	.034	.110**	.131**	.045	-.011	.184**	.055	c	-.007
	Sig. (bilateral)	.000	.000		.000	.000	.549	.411	.000	.000	.000	.210	.000	.297	.001	.000	.168	.739	.000	.097	.	.833
	N	924	924	924	924	924	924	924	924	924	924	924	924	924	924	924	924	924	924	924	924	924
Crit4	Correlación de Pearson	.078*	.066*	.253**	1	.169**	-.008	.046	.150**	.201**	.107**	.092**	.126**	.074*	.220**	.253**	.042	-.004	.201**	.144**	c	-.003
	Sig. (bilateral)	.018	.045	.000		.000	.818	.164	.000	.000	.001	.005	.000	.024	.000	.000	.201	.898	.000	.000	.	.936
	N	924	924	924	924	924	924	924	924	924	924	924	924	924	924	924	924	924	924	924	924	924
Crit5	Correlación de Pearson	.260**	.225**	.258**	.169**	1	-.015	.066*	.230**	.257**	.162**	.111**	.125**	-.025	.042	.121**	-.038	-.008	.159**	.055	c	-.005
	Sig. (bilateral)	.000	.000	.000	.000		.658	.046	.000	.000	.000	.001	.000	.446	.206	.000	.251	.806	.000	.096	.	.877
	N	924	924	924	924	924	924	924	924	924	924	924	924	924	924	924	924	924	924	924	924	924
Crit6	Correlación de Pearson	-.029	-.033	-.020	-.008	-.015	1	.097**	.074*	.021	.011	.011	-.019	.198**	-.023	-.020	-.019	-.010	-.006	-.034	c	-.006
	Sig. (bilateral)	.373	.311	.549	.818	.658		.003	.024	.517	.744	.742	.559	.000	.493	.549	.569	.766	.852	.301	.	.851
	N	924	924	924	924	924	924	924	924	924	924	924	924	924	924	924	924	924	924	924	924	924
Crit7	Correlación de Pearson	-.020	.042	.027	.046	.066*	.097**	1	.330**	.224**	.415**	.192**	.030	.094**	.055	.027	-.017	-.024	.047	.008	c	-.015
	Sig. (bilateral)	.546	.203	.411	.164	.046	.003		.000	.000	.000	.000	.359	.004	.093	.411	.603	.465	.155	.813	.	.644
	N	924	924	924	924	924	924	924	924	924	924	924	924	924	924	924	924	924	924	924	924	924
Crit8	Correlación de Pearson	.100**	.095**	.187**	.150**	.230**	.074*	.330**	1	.497**	.521**	.295**	.095**	.072*	.042	.059	-.009	-.018	.036	.014	c	-.011
	Sig. (bilateral)	.002	.004	.000	.000	.000	.024	.000		.000	.000	.000	.004	.029	.197	.072	.776	.581	.272	.668	.	.727
	N	924	924	924	924	924	924	924	924	924	924	924	924	924	924	924	924	924	924	924	924	924
Crit9	Correlación de Pearson	.102**	.182**	.261**	.201**	.257**	.021	.224**	.497**	1	.417**	.275**	.185**	.068*	.114**	.137**	.033	-.014	.130**	.127**	c	.121**
	Sig. (bilateral)	.002	.000	.000	.000	.000	.517	.000	.000		.000	.000	.000	.039	.001	.000	.310	.677	.000	.000	.	.000
	N	924	924	924	924	924	924	924	924	924	924	924	924	924	924	924	924	924	924	924	924	924
Crit10	Correlación de Pearson	.063	.053	.148**	.107**	.162**	.011	.415**	.521**	.417**	1	.308**	.053	.089**	.030	.049	.002	-.025	.028	.049	c	.062*
	Sig. (bilateral)	.057	.110	.000	.001	.000	.744	.000	.000	.000		.000	.108	.007	.361	.136	.961	.454	.393	.138	.	.060
	N	924	924	924	924	924	924	924	924	924	924	924	924	924	924	924	924	924	924	924	924	924
Crit11	Correlación de Pearson	.197**	.152**	.041	.092**	.111**	.011	.192**	.295**	.275**	.308**	1	.081*	.081*	.056	.041	.024	-.046	.027	-.007	c	.023
	Sig. (bilateral)	.000	.000	.210	.005	.001	.742	.000	.000	.000	.000		.014	.013	.089	.210	.467	.167	.408	.823	.	.479
	N	924	924	924	924	924	924	924	924	924	924	924	924	924	924	924	924	924	924	924	924	924
Crit12	Correlación de Pearson	.040	.060	.136**	.126**	.125**	-.019	.030	.095**	.185**	.053	.081*	1	.142**	.068*	.031	.074*	.093**	.104**	.121**	c	-.007
	Sig. (bilateral)	.220	.067	.000	.000	.000	.559	.359	.004	.000	.108	.014		.000	.040	.349	.025	.005	.002	.000	.	.838
	N	924	924	924	924	924	924	924	924	924	924	924	924	924	924	924	924	924	924	924	924	924
Crit13	Correlación de Pearson	.020	.006	.034	.074*	-.025	.198**	.094**	.072*	.068*	.089**	.081*	.142**	1	.051	.034	.074*	.131**	.030	.065*	c	-.011
	Sig. (bilateral)	.538	.867	.297	.024	.446	.000	.004	.029	.039	.007	.013	.000		.119	.297	.024	.000	.368	.047	.	.746
	N	924	924	924	924	924	924	924	924	924	924	924	924	924	924	924	924	924	924	924	924	924
Crit14	Correlación de Pearson	-.007	-.015	.110**	.220**	.042	-.023	.055	.042	.114**	.030	.056	.068*	.051	1	.424**	-.016	-.013	.176**	.011	c	.133**
	Sig. (bilateral)	.843	.646	.001	.000	.206	.493	.093	.197	.001	.361	.089	.040	.119		.000	.625	.703	.000	.744	.	.000
	N	924	924	924	924	924	924	924	924	924	924	924	924	924	924	924	924	924	924	924	924	924
Crit15	Correlación de Pearson	.038	.057	.131**	.253**	.121**	-.020	.027	.059	.137**	.049	.041	.031	.034	.424**	1	.021	-.011	.241**	.086**	c	-.007
	Sig. (bilateral)	.252	.084	.000	.000	.000	.549	.411	.072	.000	.136	.210	.349	.297	.000		.519	.739	.000	.009	.	.833
	N	924	924	924	924	924	924	924	924	924	924	924	924	924	924	924	924	924	924	924	924	924
Crit16	Correlación de Pearson	-.026	-.027	.045*	.042*	-.038*	-.019	-.017	-.009	.033	.002	-.024	.074*	.131**	-.016	.021	1	.023	.263**	.204**	c	-.016
	Sig. (bilateral)	.429	.408	.168	.201	.251	.569	.603	.776	.310	.961	.467	.025	.000	.625	.519		.494	.000	.000	.	.626
	N	924	924	924	924	924	924	924	924	924	924	924	924	924	924	924	924	924	924	924	924	924
Crit17	Correlación de Pearson	-.016	-.019	-.011	-.004	-.008	-.010	-.024	-.018	-.014	-.025	-.046	.093**	.119**	-.013	-.011	.023	1	.091**	.042	c	-.003
	Sig. (bilateral)	.621	.573	.739	.898	.806	.766	.465	.581	.677	.464	.167	.005	.000	.703	.739	.494		.005	.198	.	.917
	N	924	924	924	924	924	924	924	924	924	924	924	924	924	924	924	924	924	924	924	924	924
Crit18	Correlación de Pearson	.055	.050	.184**	.201**	.159**	-.006	.047	.036	.130**	.028	.027	.104**	.030	.176**	.241**	.263**	.091**	1	.304**	c	.075*
	Sig. (bilateral)	.098	.126	.000	.000	.000	.852	.155	.272	.000	.393	.408	.002	.368	.000	.000	.000	.005		.000	.	.022
	N	924	924	924	924	924	924	924	924	924	924	924	924	924	924	924	924	924	924	924	924	924
Crit19	Correlación de Pearson	-.014	-.026	.055	.144**	.055	-.034	.008	.014	.127**	.049	-.007	.121**	.065*	.011	.086**	.204**	.042	.304**	1	c	.085**
	Sig. (bilateral)	.672	.421	.097	.000	.096	.301	.813	.668	.000	.138	.823	.000	.047	.744	.009	.000	.198	.000		.	.010
	N	924	924	924	924	924	924	924	924	924	924	924	924	924	924	924	924	924	924	924	924	924
Crit20	Correlación de Pearson	c	c	c	c	c	c	c	c	c	c	c	c	c	c	c	c	c	c	c	c	c
	Sig. (bilateral)	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.
	N	924	924	924	924	924	924	924	924	924	924	924	924	924	924	924	924	924	924	924	924	924
Crit21	Correlación de Pearson	-.010	-.012	-.007	-.003	-.005	-.006	-.015	-.011	.121**	.062*	.023	-.007	-.011	.133**	-.007	-.016	-.003	.075*	.085**	c	1
	Sig. (bilateral)	.755	.722	.833	.936	.877	.851	.644	.727	.000	.060	.479	.838	.746	.000	.833	.626	.917	.022	.010	.	
	N	924	924	924	924	924	924	924	924	924	924	924	924	924	924	924	924	924	924	924	924	924

**. La correlación es significativa al nivel 0,01 (bilateral).

*. La correlación es significante al nivel 0,05 (bilateral).

c. No se puede calcular porque al menos una variable es constante.

Anexo 11. Análisis de confiabilidad del instrumento preliminar

Estadísticos de fiabilidad

Alfa de Cronbach	Alfa de Cronbach basada en los elementos tipificados	N de elementos
,615	,651	20

Estadísticos total-elemento

	Media de la escala si se elimina el elemento	Varianza de la escala si se elimina el elemento	Correlación elemento-total corregida	Alfa de Cronbach si se elimina el elemento
C1	34,73	8,686	,151	,610
C2	34,78	8,585	,162	,609
C3	34,67	8,581	,245	,599
C4	34,59	8,867	,396	,601
C5	34,62	8,889	,263	,604
C6	34,59	9,161	,074	,615
C7	34,92	7,778	,294	,589
C8	34,75	7,947	,434	,572
C9	34,68	8,296	,407	,582
C10	34,90	7,557	,410	,568
C11	35,30	8,639	,033	,641
C12	34,61	8,916	,192	,607
C13	34,74	8,343	,230	,600
C14	34,69	8,614	,203	,603
C15	34,65	8,658	,279	,598
C16	34,95	7,860	,218	,607
C17	34,60	9,107	,085	,614
C18	34,88	7,750	,317	,585
C19	34,82	8,256	,170	,612
C21	34,59	9,214	-,011	,617

Anexo 12. Prueba T para los criterios del instrumento preliminar

Prueba de muestras independientes

		Prueba T para la igualdad de medias		
		Sig. (bilateral)	Diferencia de medias	Error típ. de la diferencia
C1	Se han asumido varianzas iguales	,000	-,233	,053
	No se han asumido varianzas iguales	,000	-,233	,054
C2	Se han asumido varianzas iguales	,000	-,260	,070
	No se han asumido varianzas iguales	,000	-,260	,071
C3	Se han asumido varianzas iguales	,001	-,174	,051
	No se han asumido varianzas iguales	,001	-,174	,053
C4	Se han asumido varianzas iguales	,092	-,047	,027
	No se han asumido varianzas iguales	,103	-,047	,028
C5	Se han asumido varianzas iguales	,001	-,105	,032
	No se han asumido varianzas iguales	,002	-,105	,033
C6	Se han asumido varianzas iguales	,145	-,023	,016
	No se han asumido varianzas iguales	,159	-,023	,016
C7	Se han asumido varianzas iguales	,000	-,838	,080
	No se han asumido varianzas iguales	,000	-,838	,083
C8	Se han asumido varianzas iguales	,000	-,500	,066
	No se han asumido varianzas iguales	,000	-,500	,068
C9	Se han asumido varianzas iguales	,000	-,314	,056
	No se han asumido varianzas iguales	,000	-,314	,058
C10	Se han asumido varianzas iguales	,000	-,815	,073
	No se han asumido varianzas iguales	,000	-,815	,075
C11	Se han asumido varianzas iguales	,000	-,466	,089
	No se han asumido varianzas iguales	,000	-,466	,090
C12	Se han asumido varianzas iguales	,049	-,070	,035
	No se han asumido varianzas iguales	,057	-,070	,036
C13	Se han asumido varianzas iguales	,000	-,454	,081
	No se han asumido varianzas iguales	,000	-,454	,083

Prueba de muestras independientes

		Prueba T para la igualdad de medias		
		Sig. (bilateral)	Diferencia de medias	Error típ. de la diferencia
C14	Se han asumido varianzas iguales	,001	-,174	,051
	No se han asumido varianzas iguales	,001	-,174	,053
C15	Se han asumido varianzas iguales	,001	-,151	,044
	No se han asumido varianzas iguales	,001	-,151	,045
C16	Se han asumido varianzas iguales	,000	-,733	,090
	No se han asumido varianzas iguales	,000	-,733	,093
C17	Se han asumido varianzas iguales	,073	-,035	,019
	No se han asumido varianzas iguales	,083	-,035	,020
C18	Se han asumido varianzas iguales	,000	-,709	,085
	No se han asumido varianzas iguales	,000	-,709	,087
C19	Se han asumido varianzas iguales	,000	-,558	,086
	No se han asumido varianzas iguales	,000	-,558	,088
C21	Se han asumido varianzas iguales	,968	-,001	,016
	No se han asumido varianzas iguales	,968	-,001	,016

Anexo 13. Análisis factorial del instrumento para evaluar reactivos de opción múltiple

KMO y prueba de Bartlett

Medida de adecuación muestral de Kaiser-Meyer-Olkin.		,666
Prueba de esfericidad de Bartlett	Chi-cuadrado aproximado	599,285
	gl	91
	Sig.	,000

Varianza total explicada

	Autovalores iniciales			Sumas de las saturaciones al cuadrado de la extracción	
Componente	Total	% de la varianza	% acumulado	Total	% de la varianza
1	2,696	19,256	19,256	2,696	19,256
2	1,802	12,874	32,130	1,802	12,874
3	1,350	9,643	41,773	1,350	9,643
4	1,149	8,206	49,979	1,149	8,206
5	1,062	7,582	57,561	1,062	7,582
6	,913	6,523	64,084		
7	,851	6,075	70,160		
8	,756	5,403	75,563		
9	,704	5,028	80,591		
10	,681	4,862	85,453		
11	,613	4,380	89,833		
12	,549	3,920	93,753		
13	,539	3,853	97,606		
14	,335	2,394	100,000		

Matriz de estructura

	Componente				
	1	2	3	4	5
C1		,849			-,101
C2	,112	,865	,137		
C3	,307	,322	,335	,154	-,406
C5	,338	,411	,241		-,487
C7	,656	-,123		,202	,303
C8	,738	,111	,170	,130	
C9	,655	,283	,256		
C10	,721			,261	
C13	,314	,162	,139		,773
C14	,121		,814		
C15		,154	,783	,211	
C16	,130			,711	,118
C18	,249		,343	,608	
C19	,105			,642	-,205

Método de extracción: Análisis de componentes principales.
Metodo de rotación: Normalización Oblimin con Kaiser.

Estadísticos de fiabilidad

Alfa de Cronbach	Alfa de Cronbach basada en los elementos tipificados	N de elementos
,627	,659	14

Anexo 14. Clasificación de reactivos de acuerdo con Downing (2009) y Haladyna (2013)

Item ID	P	Rpbis	Nivel de acuerdo con Downing	Tipo de acuerdo con Haladyna
P101	0.89	0.25	II	1
P102	0.77	0.17	II	1
P103	0.84	0.14	NA	2
P104	0.71	0.25	I	1
P105	0.77	0.2	II	1
P106	0.71	0.25	I	1
P107	0.82	0.24	II	1
P108	0.9	0.13	NA	2
P109	0.72	0.23	I	1
P110	0.3	-0.17	NA	5
P111	0.63	0.31	I	1
P112	0.54	0.24	I	4
P113	0.69	0.33	I	1
P114	0.61	0.21	I	1
P115	0.5	0.2	I	4
P116	0.57	0.26	I	4
P117	0.7	0.31	I	1
P118	0.88	0.13	NA	2
P119	0.79	0.07	NA	2
P120	0.77	0.15	II	1
P121	0.68	0.33	I	1
P122	0.6	0.35	I	1
P123	0.76	0.35	II	1
P124	0.66	0.43	I	1
P125	0.31	-0.02	NA	5
P126	0.75	0.33	I	1
P127	0.51	0.16	NA	4
P128	0.36	0.3	III	4
P129	0.52	0.24	I	4
P130	0.54	0.36	I	4
P131	0.81	0.33	II	1
P132	0.77	0.25	II	1

Item ID	P	Rpbis	Nivel de acuerdo con Downing	Tipo de acuerdo con Haladyna
P133	0.51	0.37	I	4
P134	0.52	0.36	I	4
P135	0.79	0.15	II	1
P136	0.56	0.08	NA	5
P137	0.65	0.35	I	1
P138	0.88	0.18	II	1
P139	0.67	0.24	I	1
P140	0.58	0.35	I	4
P141	0.61	0.35	I	1
P142	0.68	0.39	I	1
P143	0.78	0.21	II	1
P144	0.68	0.36	I	1
P145	0.72	0.19	NA	1
P146	0.48	0.25	I	4
P147	0.58	0.32	I	4
P148	0.78	0.34	II	1
P149	0.62	0.33	I	1
P150	0.31	0.14	III	5
P151	0.86	0.28	II	1
P152	0.68	0.31	I	1
P153	0.53	0.32	I	4
P154	0.56	0.14	NA	5
P155	0.62	0.36	I	1
P156	0.38	0.37	III	4
P157	0.63	0.41	I	1
P158	0.77	0.27	II	1
P159	0.41	0.3	III	4
P160	0.41	0.27	III	4

Item ID	P	Rpbis	Nivel de acuerdo con Downing	Tipo de acuerdo con Haladyna
P201	0.69	0.15	NA	1
P202	0.72	0.18	NA	1
P203	0.68	0.3	I	1
P204	0.82	0.3	II	1
P205	0.52	0.17	NA	4
P206	0.7	0.39	I	1
P207	0.5	0.3	I	4
P208	0.55	0.27	I	4
P209	0.52	0.32	I	4
P210	0.85	0.24	II	1
P211	0.81	0.25	II	1
P212	0.67	0.45	I	1
P213	0.58	0.33	I	4
P214	0.3	0.18	III	4
P215	0.49	0.25	I	4
P216	0.89	0.16	II	1
P217	0.33	0.1	III	5
P218	0.76	0.01	NA	2
P219	0.62	0.49	I	1
P220	0.74	0.33	I	1
P221	0.64	0.28	I	1
P222	0.8	0.36	II	1
P223	0.37	0.29	III	4
P224	0.35	0.17	III	4
P225	0.81	0.35	II	1
P226	0.68	0.29	I	1
P227	0.68	0.37	I	1
P228	0.1	-0.16	IV	5
P229	0.47	0.34	I	4
P230	0.69	0.48	I	1
P231	0.65	0.33	I	1
P232	0.72	0.49	I	1
P233	0.63	0.37	I	1
P234	0.5	0.16	NA	4
P235	0.52	0.41	I	4
P236	0.6	0.5	I	1
P237	0.46	0.34	I	4

Item ID	P	Rpbis	Nivel de acuerdo con Downing	Tipo de acuerdo con Haladyna
P238	0.39	0.4	III	4
P239	0.48	0.34	I	4
P240	0.33	0.49	III	4
P241	0.41	0.33	III	4
P242	0.59	0.42	I	4
P243	0.54	0.39	I	4
P244	0.93	0.21	IV	3
P245	0.64	0.23	I	1
P246	0.44	0.33	III	4
P247	0.62	0.38	I	1
P248	0.64	0.07	NA	2
P249	0.73	0.43	I	1
P250	0.31	0.34	III	4
P251	0.73	0.35	I	1
P252	0.1	-0.2	IV	5
P253	0.69	0.25	I	1
P254	0.79	0.05	NA	2
P255	0.8	0.15	II	1
P256	0.55	0.24	I	4
P257	0.65	0.2	I	1
P258	0.38	0.36	III	4
P259	0.74	0.44	I	1
P260	0.2	-0.02	IV	5
P261	0.34	0.32	III	4
P262	0.79	0.42	II	1
P263	0.52	0.45	I	4
P264	0.54	0.33	I	4
P265	0.62	0.41	I	1
P266	0.42	0.17	III	4
P267	0.5	0.34	I	4
P268	0.73	0.24	I	1
P269	0.58	0.18	NA	4
P270	0.68	0.11	NA	2
P301	0.39	0.32	III	4
P302	0.65	0.42	I	1

Item ID	P	Rpbis	Nivel de acuerdo con Downing	Tipo de acuerdo con Haladyna
P303	0.66	0.46	I	1
P304	0.79	0.29	II	1
P305	0.37	0.32	III	4
P306	0.64	0.32	I	1
P307	0.63	0.41	I	1
P308	0.78	0.19	II	1
P309	0.47	0.31	I	4
P310	0.6	0.35	I	1
P311	0.96	0.09	IV	3
P312	0.49	0.4	I	4
P313	0.67	0.25	I	1
P314	0.58	0.38	I	4
P315	0.53	0.49	I	4
P316	0.44	0.37	III	4
P317	0.69	0.29	I	1
P318	0.65	0.38	I	1
P319	0.75	0.39	I	1
P320	0.73	0.24	I	1
P321	0.5	0.44	I	4
P322	0.57	0.42	I	4
P323	0.48	0.22	I	4
P324	0.73	0.23	I	1
P325	0.38	0.12	III	5
P326	0.52	0.47	I	4
P327	0.53	0.3	I	4
P328	0.71	0.39	I	1
P329	0.75	0.33	I	1
P330	0.68	0.34	I	1
P331	0.7	0.3	I	1
P332	0.28	0.3	III	4
P333	0.54	0.23	I	4
P334	0.48	0.47	I	4
P335	0.29	0.36	III	4
P336	0.49	0.41	I	4
P337	0.7	0.37	I	1
P338	0.54	0.22	I	4
P339	0.46	0.43	I	4

Item ID	P	Rpbis	Nivel de acuerdo con Downing	Tipo de acuerdo con Haladyna
P340	0.47	0.47	I	4
P341	0.41	0.43	III	4
P342	0.66	0.37	I	1
P343	0.58	0.36	I	4
P344	0.41	0.45	III	4
P345	0.33	0.29	III	4
P346	0.48	0.45	I	4
P347	0.51	0.18	NA	4
P348	0.36	0.34	III	4
P349	0.63	0.21	I	1
P350	0.6	0.19	NA	1
P351	0.51	0.03	NA	5
P352	0.74	0.32	I	1
P353	0.96	0.09	IV	3
P354	0.41	0.33	III	4
P355	0.12	0.14	IV	5
P356	0.57	0.45	I	4
P357	0.64	0.18	NA	1
P358	0.54	0.27	I	4
P359	0.46	0.51	I	4
P360	0.75	0.29	I	1
P361	0.83	0.25	II	1
P362	0.19	0.53	IV	4
P363	0.47	0.51	I	4
P364	0.78	0.28	II	1
P365	0.35	0.5	III	4
P366	0.77	0.21	II	1
P367	0.48	0.55	I	4
P368	0.46	0.29	I	4
P369	0.74	0.25	I	1
P370	0.83	0.18	II	1

Anexo 15. Correlaciones entre el número de distractores no funcionales (NFD), dificultad y discriminación de los reactivos

Correlations

				NFD	Dificultad	Discriminacion
NFD	Pearson Correlation			1	.506**	-.253**
	Sig. (2-tailed)				.000	.000
	N			200	200	200
	Bootstrap[c]	Bias		0	-.004	.001
		Std. Error		0	.056	.067
		95% Confidence Interval	Lower	1	.384	-.379
			Upper	1	.600	-.120
Dificultad	Pearson Correlation			.506**	1	-.014
	Sig. (2-tailed)			.000		.842
	N			200	200	200
	Bootstrap[c]	Bias		-.004	0	-.005
		Std. Error		.056	0	.112
		95% Confidence Interval	Lower	.384	1	-.243
			Upper	.600	1	.196
Discriminacion	Pearson Correlation			-.253**	-.014	1
	Sig. (2-tailed)			.000	.842	
	N			200	200	200
	Bootstrap[c]	Bias		.001	-.005	0
		Std. Error		.067	.112	0
		95% Confidence Interval	Lower	-.379	-.243	1
			Upper	-.120	.196	1

**. Correlation is significant at the 0.01 level (2-tailed).

Anexo 16. Prueba de los rangos con signo de Wilcoxon para comparar la dificultad (Dif) y discriminación (Disc) entre reactivos estándar (ST) y reactivos con errores (FL)

Resumen de prueba de hipótesis

	Hipótesis nula	Test	Sig.	Decisión
1	La mediana de las diferencias entre STDif y FLDif es igual a 0.	Prueba de Wilcoxon de los rangos con signo de muestras relacionadas	,085	Retener la hipótesis nula.

Se muestran las significancias asintóticas. El nivel de significancia es ,05.

Resumen de prueba de hipótesis

	Hipótesis nula	Test	Sig.	Decisión
1	La mediana de las diferencias entre STDisc y FLDisc es igual a 0.	Prueba de Wilcoxon de los rangos con signo de muestras relacionadas	,226	Retener la hipótesis nula.

Se muestran las significancias asintóticas. El nivel de significancia es ,05.

Anexo 17. Correlación entre número de errores y parámetros psicométricos

Correlations

				Errores	P	Rpbis
Errores	Pearson Correlation			1	.042	-.010
	Sig. (2-tailed)				.556	.884
	N			200	200	200
	Bootstrap[c]	Bias		0	-.002	-.001
		Std. Error		0	.064	.076
		95% Confidence Interval	Lower	1	-.089	-.153
			Upper	1	.161	.145
P	Pearson Correlation			.042	1	-.014
	Sig. (2-tailed)			.556		.842
	N			200	200	200
	Bootstrap[c]	Bias		-.002	0	-.005
		Std. Error		.064	0	.109
		95% Confidence Interval	Lower	-.089	1	-.242
			Upper	.161	1	.186
Rpbis	Pearson Correlation			-.010	-.014	1
	Sig. (2-tailed)			.884	.842	
	N			200	200	200
	Bootstrap[c]	Bias		-.001	-.005	0
		Std. Error		.076	.109	0
		95% Confidence Interval	Lower	-.153	-.242	1
			Upper	.145	.186	1

Anexo 18. Ejemplo de la base de datos de las respuestas de los estudiantes en los exámenes evaluados.

305346346														
305580728	D	B	A	D	B	A	C	B	A	D	C	A	C	B
306012853	D	B	A	A	B	B	B	B	A	D	D	A	B	D
306040133	D	A	A	D	B	A	B	B	A	B	D	D	B	B
306044911	*	*	*	*	*	*	*	*	*	*	*	*	*	*
306052927														
306058936														
306065031	A	B	B	D	A	A	B	B	A	B	B	A	C	B
306070699	D	A	A	A	A	A	A	B	A	B	C	B	B	C
306110502	D	B	A	A	D	B	B	B	A	B	B	C	C	A
306113149	D	B	A	D	B	A	B	B	A	D	B	D	B	B
306114366	D	B	A	D	B	A	B	B	A	D	B	A	B	C
306140288	D	D	C	B	B	A	B	B	D	B	B	B	B	B
306143447	C	B	C	D	A	D	B	B	D	D	D	A	C	B
306151800	B	B	A	D	C	A	D	B	A	D	B	A	B	A
306152890	D	B	A	D	B	B	B	A	A	D	B	D	B	C
306180273	A	A	A	D	A	B	D	B	A	B	C	B	D	A
306183982	B	B	B	D	A	A	B	B	A	D	C	B	C	A
306207639	D	D	A	A	B	D	B	B	D	A	D	B	D	B
306211867	d	B	B	D	B	B	b	B	A	b	D	d	B	A
306226346														
306245606	D	A	A	A	A	B	B	B	A	D	D	D	C	B
306251436														
306264409	C	B	A	A	B	B	D	B	A	D	A	D	B	D
306312249														