Revista de Estudios y Experiencias en Educación
Vol. 18 Nº 37, agosto, 2019 pp. 143 - 151

Análisis de los instrumentos evaluativos de la asignatura Bases Biológicas y Neurobiológicas del Desarrollo

Félix Fernando Aragón[a] y Carlos Albaca Paraván*[b]
Universidad Nacional de Tucumán, Facultad de Medicina, San Miguel de Tucumán, Argentina.

Recibido: 16 febrero 2019
Aceptado: 22 abril 2019

RESUMEN. Incluso después de haber sido administrado un examen, no siempre es seguro que haya funcionado como debería. Por ello, en el proceso de evaluación del aprendizaje, es crucial el establecimiento de la calidad de los instrumentos con los que se lleva a cabo esta tarea. En este contexto, los docentes de la asignatura Bases Biológicas y Neurobiológicas del Desarrollo, en el año 2018, iniciaron un proceso de análisis de la calidad de los elementos que forman parte de los exámenes parciales aplicando características psicométricas a los mismos. Este trabajo muestra y analiza los resultados obtenidos de dicha evaluación y sugiere modificaciones que permitan mejorar la calidad de los exámenes parciales de la asignatura.

PALABRAS CLAVE. Calidad educativa; Evaluación; Psicometría.

Analysis of the evaluation instruments of the subject Biological and Neurobiological Bases of Development

ABSTRACT. Even after an exam has been administered, it is not always sure that it worked as it should. Therefore, in the evaluation process of learning, it is crucial to establish the quality of the instruments with which this task is carried out. In this context, the teachers of the subject Biological Bases and Neurobiological Development, in 2018, began a process of analysis of the quality of the items that are part of the partial examinations applying psychometric characteristics to them. This work shows the results of the analyses obtained from this evaluation and suggests modifications that improve the quality of the partial exams of the subject.

KEYWORDS. Educational quality; Evaluation; Psychometrics.

1. INTRODUCCIÓN Y MARCO TEÓRICO

La evaluación del aprendizaje ha sido siempre asunto de investigación y estudio ya que tiene como función valorar las habilidades y conocimientos obtenidos por los alumnos, es así que a lo largo de los años, se han implementado diversos tipos de instrumentos evaluativos los cuales contienen reactivos o ítems que se confeccionan de acuerdo al propósito de la medición (Beltrán Martínez, Márquez y Hernández, 2015).

*Correspondencia: Carlos Albaca Paraván. Dirección: Crisóstomo Álvarez Nº 809, San Miguel de Tucumán, Argentina. Correos Electrónicos: felix.aragon82@gmail.com[a], calbaca@herrera.unt.edu.ar[b].

doi: 10.21703/rexe.20191837aragon4 *ISSN 0717-6945 ISSN en línea 0718-5162*

Córdova Islas (2010) define la evaluación como el procedimiento sistemático y comprensivo en el cual se utilizan múltiples estrategias, tales como: cuestionarios, inventarios, entrevistas, exámenes normalizados o de criterio, exámenes orales, pruebas cortas, portafolios, presentaciones, etc. La evaluación es, entonces, un conjunto de estrategias consignadas a mejorar la calidad de la enseñanza.

Mediante la evaluación se pueden obtener respuestas a diferentes interrogantes: ¿Qué deben aprender los estudiantes? ¿Hasta qué punto lo están aprendiendo? ¿Están aprendiendo lo que estamos enseñando? ¿Cómo podemos mejorar el proceso de enseñanza-aprendizaje?

Por su lado, González Pérez (2005) reafirma lo mencionado por Córdova Islas (2010) enunciando que la evaluación desempeña diversas funciones como:

- Definir los significados atribuidos a las condiciones de éxito y fracaso escolar, rendimiento educativo, buenos y malos alumnos y docentes, calidad de la enseñanza, progreso escolar, excelencia escolar.
- Certificar el saber, ya que las instituciones educativas (escuelas, colegios, universidades, institutos, etc.) otorgan títulos, diplomas o certificados a partir de los resultados de la evaluación. Por ello, socialmente se les imputa la propiedad de simbolizar la posesión del saber y la competencia.
- Organizar y gestionar la educación, ya que los resultados de la evaluación permiten al docente establecer las medidas de promoción, deserción, repitencia, certificación y demás aspectos que facilitan el trabajo de los establecimientos educativos y el paso de los alumnos por los diferentes niveles formativos.
- Orientar, diagnosticar, pronosticar, crear el ambiente escolar, afianzar el aprendizaje, retroalimentar, motivar, preparar a los alumnos para la vida en sociedad.

AFACIMERA (2002), por su lado, concluye que la evaluación brinda datos que permiten tomar mejores decisiones pedagógicas. Los exámenes son los instrumentos que utilizan los docentes para conseguir información del aprendizaje de sus estudiantes y luego, en base a ésta, decidir si brindarles una recuperación, promoverlos o aplazarlos. Destaca también que, evaluar mediante exámenes no es sólo una actividad escolar ya que muchos establecimientos del sector de la salud plantean y emplean exámenes y pruebas para seleccionar residentes, certificar especialistas, otorgar becas, etc.

En el área de las Ciencias de la Salud, los exámenes a estudiantes y profesionales involucran una enorme responsabilidad social ya que por medio de éstos, se faculta una práctica profesional directamente vinculada con las personas y su salud. Por ello, la construcción de exámenes que brinden resultados objetivos, válidos y confiables es todo un desafío, ya que toda la sociedad confía en que este "control de calidad" se haya llevado a cabo seriamente en los establecimientos correspondientes.

Por su lado, Gimeno Sacristán y Carbone (1992) plantean pensar a la evaluación como una justificación de la validez de las estrategias didácticas, y Palmer y Devitt (2007) completan esta idea enunciando que la evaluación sirve para buscar información que ayude a concluir si la metodología utilizada fue correcta o no, y en qué medida lo fue, de forma de dirigir un proceso de enseñanza que culmine en la obtención de los resultados de aprendizaje propuestos preliminarmente. Por lo que, desde el punto de vista de este enfoque, uno de las grandes utilidades de la evaluación es el de ser un instrumento de investigación didáctica (Córdova Islas, 2010).

La cátedra de Bases Biológicas y Neurobiológicas del Desarrollo perteneciente a la carrera de Profesorado en Educación Especial del Instituto Decroly (Tucumán, Argentina), comenzó a utilizar desde 2018 herramientas para abordar procesos de mejora de sus instrumentos de evaluación: exámenes parciales escritos de selección múltiple, siendo la primera cátedra en realizar este tipo de revisiones en la institución.

Clásicamente se señala que las mayores desventajas del tipo de exámenes de selección múltiple es que, por un lado, evalúan conocimiento de tipo memorístico más que de razonamiento elaborado, y por el otro, se centran en reconocer la respuesta correcta en vez de recuperarla de la memoria. Esto ocurre cuando los exámenes son construidos en formato de estímulo pobre de contexto, sin embargo, es posible diseñar ítems con descripciones ricas del contexto que simulan casos reales y evalúan más adecuadamente competencias. Dado el empleo tan extendido del uso de este tipo de exámenes, existe un buen grado de evidencia sobre sus características psicométricas (Durante, 2006).

La evaluación de los exámenes de selección múltiple se realizó teniendo en cuenta las características psicométricas: Índices de Dificultad de los ítems (p) (Crocker y Algina, 2008), Índice de Discriminación (ID) (Ebel y Frisbie, 1991), Norma Discriminativa (ND) y Relación Discriminativa (RD) (Díaz Rojas y Leyva Sánchez, 2013) de los elementos que componen los exámenes de la materia. Además, se realizó un análisis de distribución de los distractores funcionales (DF) (Tarrant, Ware, y Mohammed, 2009) de las preguntas de los exámenes.

Se decidió utilizar estas características psicométricas ya que son indicadores básicos que se utilizan para el análisis en exámenes objetivos (; De Los Santos Lázaro, 2010; Mercau et al., 2013; Pérez Tapia, Acuña Aguila, & Arratia Cuela, 2007), y la metodología de trabajo se llevó a cabo de forma similar a la usada por evaluaciones llevadas a cabo, por ejemplo, en el EXHCOBA de la UABC en México (Escudero, Reyna y Morales, 2000), la Especialidad de Medicina General Integral de la Medical University of Santiago de Cuba (Díaz Rojas y Leyva Sánchez, 2013), la UAM en México (González Cuevas, 2003), y la Cátedra de Microbiología de la Facultad de Medicina de la Universidad Nacional de Tucumán en Argentina (Vece, Lepera y Tefaha, 2012), entre otras.

El objetivo del artículo es examinar los resultados de la aplicación de los índices mencionados en los reactivos que conforman los exámenes parciales de la asignatura, para de esta manera, conseguir mejorar los instrumentos de evaluación que se utilizan.

2. MATERIALES Y MÉTODOS

Se realizó un estudio cuantitativo retrospectivo de tipo observacional para el cual se utilizaron como insumo las respuestas de los dos parciales de selección múltiple de los 30 alumnos que cursaron la materia durante el año 2018.

La asignatura corresponde al primer año del Profesorado en Educación Especial y sus objetivos resumidos son que el alumno incorpore conocimientos sobre los aparatos y sistemas del organismo humano, y conceptos relacionados:

- Sistema inmunológico y el calendario nacional de vacunas.

- Aparato reproductor masculino y femenino, fecundación, desarrollo embrionario, embarazo y parto, patologías y discapacidades producidas en estos procesos.

- Sistema nervioso (anatomía y fisiología) y la importancia del estímulo-respuesta como experiencia que ayuda a la maduración del mismo. Sustancias tóxicas y adictivas que afectan al sis-

tema nervioso con dependencia física y psíquica, y que produzcan daños en el sistema nervioso central. Prevención de acciones y comprensión de las consecuencias del uso de tóxicos.

- Órganos de los sentidos (anatomía y fisiología), con especial énfasis en la vista y oído, para comprender posteriormente patologías neurosensoriales.

La metodología de corrección de los parciales fue realizada de forma manual por los profesores de la asignatura según una clave de respuestas realizada anticipadamente.

Cada uno de los 2 parciales contó con 20 preguntas de 4 opciones cada una, una correcta y 3 distractores. Las preguntas tenían un puntaje preestablecido de 5 puntos cada una, y a cada respuesta se le asignó ese puntaje, si el alumno eligió correctamente la respuesta, o cero en caso contrario. Para dar por aprobado un examen se requirió un puntaje mínimo de 40/100.

La información obtenida de los parciales se reprodujeron a una planilla de cálculos y se calcularon p, ID, ND, RD y los porcentajes de distribución de respuesta de cada pregunta.

Se entiende como Índice de Dificultad (p) al número de examinados que aciertan a un ítem o reactivo, entre el total que intentó resolverlo (tanto por ciento si multiplicamos por 100). Este término favorece a confusiones, ya que un valor más alto indica un reactivo más fácil (mayor proporción de aciertos), no más difícil, quizás debería nombrarse índice de facilidad como sugiere Morales (2009). El grado de facilidad o dificultad de un reactivo puede precisarse en función del criterio de interpretación indicado en la Tabla 1, sugiriéndose elaborar instrumentos evaluativos que adopten una distribución con valores de p según se indica en la Tabla 2.

Tabla 1. Criterio de Interpretación de p (AFACIMERA, 2002).

Evaluación del ítem	p
Muy Fácil (MF)	[0,85 ; 1]
Relativamente Fácil (RF)	[0,69 ; 0,84]
Dificultad Media (DM)	[0,32 ; 0,68]
Relativamente Difícil (RD)	[0,16 ; 0,31]
Muy Difícil (MD)	[0 ; 0,15]

Tabla 2. Distribución de p (Escudero, Reyna, & Morales, 2000).

p	% de ítems en la evaluación
Muy Fácil	5%
Relativamente Fácil	20%
Dificultad Media	50%
Relativamente Difícil	20%
Muy Difícil	5%

Si el examen y un reactivo evalúan la misma habilidad, competencia o conocimiento, puede esperarse que quien obtuvo una nota alta en el examen tendrá altas probabilidades de responder acertadamente un ítem. También se puede esperar lo inverso, es decir, que quien obtuvo baja puntuación en el examen, tendrá pocas probabilidades de contestar correctamente un ítem. Así, un buen ítem/reactivo debe discriminar entre los que lograron buenas calificaciones y los que lograron bajas calificaciones en el examen. Para poder determinar esto, se puede usar el Índice

de Discriminación (ID) según el criterio mostrado en la Tabla 3, intentando generar exámenes donde se logre una distribución de las preguntas con valores de Índice de Discriminación como indica la Tabla 4.

Tabla 3. Interpretación de ID (AFACIMERA, 2002).

Calidad	ID %	Recomendación
Muy Buena	ID ≥ 40	Se debe conservar
Buena	30 ≥ ID ≥ 39	Se puede mejorar
Regular	20 ≥ ID ≥ 29	Se debe mejorar
Deficiente	ID ≤ 19	Se debe descartar o revisar

Tabla 4. Distribución de ID (AFACIMERA, 2002).

Calidad	% de ítems en la evaluación
Muy Buena	Mayor o igual a 25%
Buena	Entre 16% y 24%
Regular	Menor a 15%
Deficiente	Menor a 5%

Debe considerarse que los reactivos muy fáciles o muy difíciles no discriminan (no establecen diferencias), nos dicen que todos saben o no saben un ítem, pero no quién sabe más y quién menos. Estos ítems no contribuyen a la fiabilidad, pero eso no quiere decir que necesariamente sean malas preguntas, solo son malas discriminando. Por ello, Díaz Rojas y Leyva Sánchez (2013) entre otros, exponen usar la Norma Discriminativa (ND) que pretende calcular el valor óptimo de ID de acuerdo a su p, y con ella generar otro índice que revele si un ítem es admisible o no según su ID. Este indicador se denomina Relación Discriminativa (RD) y se lo utiliza según el criterio de la Tabla 5.

Tabla 5. Interpretación de RD (Díaz Rojas & Leyva Sánchez, 2013).

RD	Conducta a seguir
RD ≥ 1	Ítem aceptable
0,6 ≤ RD < 1	Analizar el ítem
RD ≤ 0,6	Descartar el ítem

En cuanto al análisis de los distractores, la proporción de elección de cada uno permite catalogar como DF al que fue seleccionado por al menos un 5% de los examinados. Si un distractor no cumple con esta condición quiere decir que no funcionó como tal (AFACIMERA, 2002; Case y Swanson, 2006; Guilbert, 1989; Lafourcade, 1973).

En base a todo lo anteriormente expuesto, se puede considerar que un examen es de mayor calidad a medida que:

1. La distribución real de p tiende a la distribución ideal (Tabla 2).

2. La distribución real de ID tiende a la distribución ideal (Tabla 4).

3. El número de ítems con RD < 1 tiende a cero (No se deben analizar o descartar ítems).

4. El número de ítems con máxima cantidad de DF tiende al número total de ítems del examen.

De esto se desprende que, usando este método, no se puede obtener una medición cuantitativa de la calidad, sino que la evaluación sirve de base para poder iniciar un proceso de mejora continua en la construcción de los instrumentos evaluativos que se usarán a futuro.

3. RESULTADOS

La figura 1 muestra la comparación de la distribución óptima ideal con la obtenida en cada examen de la asignatura. En el primer parcial se puede observar una pequeña tendencia a ítems con p=RD, mientras que en el segundo se puede apreciar una mayor agrupación de ítems con p=RF, consecuencia de ajustar el nivel de dificultad entre un examen y otro.

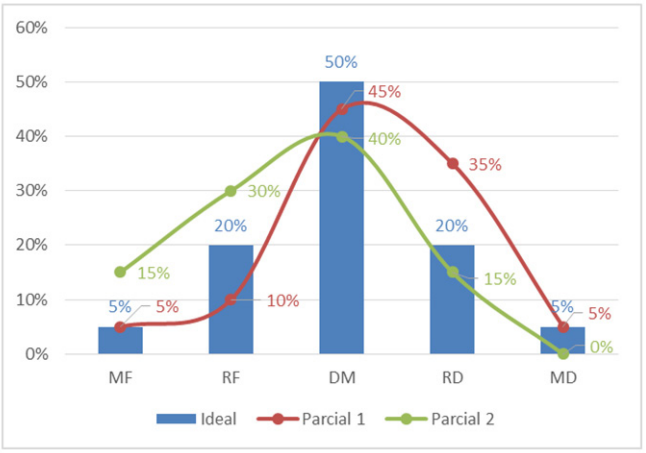

Figura 1. Distribución de p en los exámenes parciales respecto de la óptima ideal

La figura 2 muestra la distribución de ID en ambos parciales respecto de la óptima sugerida. Se puede remarcar que la cantidad de ítems con ID=MB y ID=R en ambos parciales es coherente a lo sugerido, mientras que en los otros casos estos valores están apartados de los ideales.

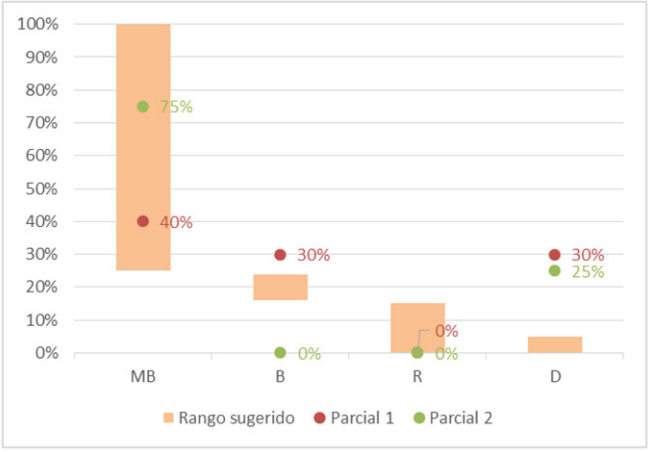

Figura 2. Distribución de ID en los dos exámenes parciales respecto a los rangos óptimos ideales

La figura 3 denota que el 5% de las preguntas del primer parcial y el 10% del segundo deben ser descartadas, mientras que el 10% de las preguntas del Parcial Nº 1 y el 5% del Parcial Nº 2 debe ser revisadas.

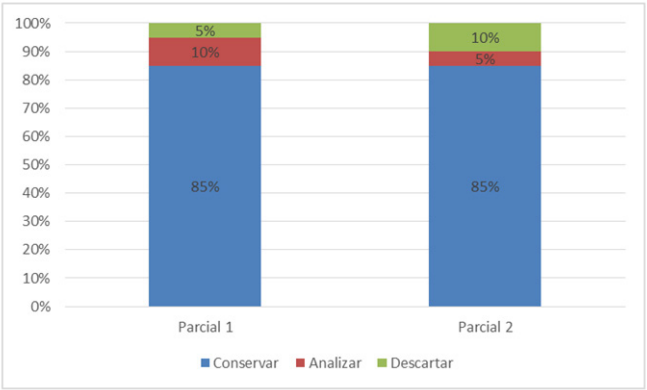

Figura 3. Distribución de RD en los dos exámenes parciales

La figura 4 muestra la distribución de distractores funcionales en cada parcial. Como cada pregunta tenía 4 opciones (donde solo una era correcta), se puede observar que en el Parcial Nº 1 todas las opciones incorrectas funcionaron como DF en el 90% de las preguntas, donde en el 10% restante una opción no funcionó como DF. Para el Parcial Nº 2, todas las opciones incorrectas funcionaron como DF en el 80% de las preguntas, donde en el 10% restante una opción no funcionó como DF y en el otro 10% 2 opciones no funcionaron como DF.

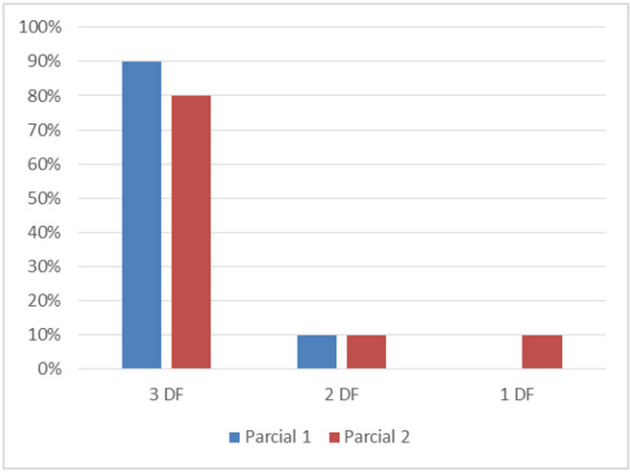

Figura 4. Distribución de los DF en los dos exámenes parciales

4. CONCLUSIONES

El presente trabajo es el inicio de un proceso sistemático de análisis de los dos exámenes parciales que utiliza la cátedra como instrumentos de evaluación a los estudiantes.

Vale recordar que, a los fines de este trabajo, la calidad de los instrumentos evaluativos está referida a: la distribución real de p, ID y RD, y la distribución del número de DF de los ítems.

En función de lo anteriormente expresado, en primer término vale mencionar que los docentes de la asignatura realizarán un esfuerzo para mejorar la distribución del Índice de Dificultad (p) en las evaluaciones, siempre tratando de alcanzar los valores óptimos sugeridos por la literatura.

En segundo término, a pesar que la RD de los ítems es adecuada, se revisarán y se reformularán las preguntas que posean ID=R y ID=D, procurando mejorar la calidad de las preguntas que conforman los exámenes parciales, proponiéndose llegar siempre a la excelencia.

En tercer término, el análisis de los instrumentos evaluativos de la asignatura será llevado a cabo periódicamente en el futuro a la espera de una mejora sustancial en los índices de calidad mencionados en este trabajo, sin dejar de lado que los mismos dependen fuertemente de cada grupo de estudiantes que realizan las evaluaciones.

Por último, con los resultados de la aplicación de estas características psicométricas a las preguntas que componen los exámenes, los docentes de la asignatura tienen como meta crear un banco de preguntas validado por estos índices de calidad, el cual sirva para crear instrumentos evaluativos cada vez mejores.

Como trabajo a futuro, se procederá a utilizar índices más complejos para calcular la dificultad de una temática para determinar los conceptos, temas o unidades que les resultó más difícil a los alumnos y poder reforzar los mismos. Además, se pretende utilizar el Coeficiente Biserial Puntual (Beltrán Martínez et al., 2015) que es una medida que se puede utilizar en el análisis para determinar la validez de un ítem con criterios externos. Finalmente se aplicará el índice de calidad (ICG) desarrollado por el Dr. Galofré (2006), teniendo en cuenta las recomendaciones para la construcción de los ítems de cada examen.

REFERENCIAS

AFACIMERA. (2002). *Evaluación Educacional* - Volumen I y II. Argentina: AFACIMERA.

Beltrán Martínez, B., Márquez, A., & Hernández, V. (2015). Diseño de un sistema de validación de reactivos con base al constructivismo. *Revista Iberoamericana de Producción Académica y Gestión Educativa*, 2(3), 1-15.

Case, S., & Swanson, D. (2006). *Como elaborar preguntas para evaluaciones escritas en el área de ciencias básicas y clínicas* (Tercera ed.). Philadelphia, Estados Unidos: Nacional Board Of Medical Examiners.

Córdova Islas, A. (2010). *Evaluación Educativa. Congreso Iberoamericano de Educación*, (pp. 1-15). Buenos Aires. Recuperado de http://webmail.adeepra.com.ar/congresos/Congreso%20IBEROAMERICANO/EVALUACION/R0009_Cordova.pdf

Crocker, L., & Algina, J. (2008). *Introduction to classical and modern test theory* (Segunda ed.). Ohio: Cengage Learning.

De Los Santos Lázaro, G. (2010). *Desarrollo, operación y evaluación de un módulo para capacitar a docentes en servicio para que mejoren sus exámenes de opción múltiple mediante el análisis gráfico de ítems*. (Tesis para obtener el grado de Maestra en Ciencias Educativas). Ensenada, Baja California, México: Universidad Autónoma de Baja California.

Díaz Rojas, P., & Leyva Sánchez, E. (2013). Metodología para determinar la calidad de los instrumentos de evaluación. *Revista Cubana de Educación Médica Superior*, 27(2), 269-286.

Durante, E. (2006). Algunos métodos de evaluación de las competencias: escalando la pirámide de Miller. *Revista del Hospital Italiano de Buenos Aires*, 26(2), 55-61.

Ebel, R., & Frisbie, D. (1991). *Essentials of educational measurement* (Quinta ed.). Michigan: Prentice Hall.

Escudero, E., Reyna, N., & Morales, M. (2000). Nivel de dificultad y poder de discriminación del Examen de Habilidades y Conocimientos Básicos (EXHCOBA). *REDIE: Revista Electrónica de Investigación Educativa*, 2(1), 12-29.

Galofré, A. (2006). *Instrucciones para calcular un índice de calidad para preguntas de selección múltiple*. Antofagasta, Chile: Universidad Católica del Norte.

Gimeno Sacristán, J., & Carbone, G. (1992). *Teoría de la enseñanza y desarrollo del currículo* (Cuarta ed.). Buenos Aires: R.E.I. Argentina.

González Cuevas, O. (2003). Evaluación de opción múltiple v.s. evaluación tradicional. Un estudio de caso en ingeniería. *Ingeniería*, 7(2), 17-37.

González Pérez, M. (2005). La evaluación del aprendizaje. *Revista Docencia Universitaria*, 6(1). Recuperado de https://revistas.uis.edu.co/index.php/revistadocencia/article/view/819.

Guilbert, J. (1989). *Guía pedagógica para el personal de la salud* (Quinta ed.). Valladolid, México: Instituto de Ciencias de la Educación.

Lafourcade, P. (1973). *Evaluación de los aprendizajes*. Buenos Aires, Argentina: Kapelusz.

Mercau, G., Coccioli, M., D'Urso, M., Siciliani, M., del Castillo, M., & Valverde, M. (2013). Impacto de la Calificación Regular en la Validación de Instrumentos de Evaluación. *Revista Facultad de Medicina*, 13(1), 37.

Morales, P. (2009). *Análisis de ítems en las pruebas objetivas*. Madrid, España: Facultad de Ciencias Humanas y Sociales de la Universidad Pontificia Comillas.

Palmer, E., & Devitt, P. (2007). Assessment of higher order cognitive skills in undergraduate education: modified essay or multiple choice questions? Research paper. *BMC Medical Education*, 7(49), 1-7. doi: https://doi.org/10.1186/1472-6920-7-49.

Pérez Tapia, J., Acuña Aguila, N., & Arratia Cuela, E. (2007). Nivel de dificultad y poder de discriminación del tercer y quinto examen parcial de la cátedra de cito-histología 2007 de la carrera de medicina de la UMSA. *Cuadernos del Hospital de Clínicas*, 53 (2),16-22.

Tarrant, M., Ware, J., & Mohammed, A. (2009). An assessment of functioning and non-functioning distractors in multiple-choice questions: a descriptive analysis. *BMC Medical Education*, 9(40), 1-8. doi: https://doi.org/10.1186/1472-6920-9-40.

Vece, M., Lepera, R., & Tefaha, L. (2012). Evaluación de calidad de exámenes de opción múltiple en microbiología aplicando diferentes índices. *Revista Argentina de Educación Médica*, 5(1), 29-34.