REICE. Revista Iberoamericana sobre Calidad, Eficacia y Cambio en Educación
2022, 20(1), 41-55.
https://doi.org/10.15366/reice2022.20.1.003

Uso del Modelo de Crédito Parcial de Rasch y Masters en la Evaluación de Competencias Matemáticas

Use of the Rasch and Masters Partial Credit Model in the Assessment of Mathematical Competencies

Eduardo Backhoff [1], * Manuel González-Montesinos [2], Yadira Pérez-Garibay [1], María Fabiana Ferreyra [1]

[1] *Métrica Educativa, A. C., México*

[2] *Centro Regional de Formación Profesional Docente de Sonora, México*

DESCRIPTORES:

Crédito parcial
Rasch
Matemáticas
Competencias
Evaluación computarizada

RESUMEN:

Recientemente, la evaluación del aprendizaje ha mejorado sustancialmente gracias al desarrollo de las ciencias cognitivas, al uso de las tecnologías digitales y a las nuevas teorías de la medición, como la Teoría de Respuestas al Ítem (TRI). Usualmente, los reactivos de un examen de selección se califican de manera dicotómica (aciertos o errores); sin embargo, los reactivos de respuesta construida se pueden calificar parcialmente. El propósito del presente estudio fue conocer la forma en que opera el modelo de crédito parcial (MCP) de Rasch y Masters (Wright y Masters, 1982). Para ello, se utilizó el Examen de Competencias Básicas de Matemáticas (examen computarizado de respuesta construida), con 547 estudiantes de ingeniería de una universidad pública mexicana. Utilizando el programa WinSteps© 4.5.5 (Linacre, 2020), se compararon las puntuaciones de los alumnos cuando sus respuestas se calificaron de manera dicotómica y de forma parcial. Los resultados muestran que ambos métodos generan puntuaciones equivalentes que, aunque no idénticas, conservan el mismo ordenamiento (correlación cercana a 1). Se concluye que el MCP de Rasch y Masters opera eficazmente y que puede utilizarse aún con reactivos cuyos elementos no implican, necesariamente, una secuencia de pasos incrementales concatenados, así como en ámbitos distintos al rendimiento académico.

KEYWORDS:

Partial credit
Rasch
Mathematics
Competencies
Computerized assessment

ABSTRACT:

Recently, the assessment of learning outcomes has improved substantially due to developments in cognitive science, in the use of digital technologies and in the application of new measurement theory, especially Item Response Theory (IRT). Usually, items in a selection exam are designed dichotomously (right or wrong) and are scored accordingly. However, items designed to be solved by constructed answers can be employed and graded with partial responses. The purpose of this study is to examine the operation of the partial credit model (PCM) developed by G. Masters, from the dichotomous Rasch Model, for this type of items (Wright & Masters, 1982). A computer administered constructed answer exam was employed: The Mathematics Basic Competencies Exam -EXCOBA Math. The exam was administered to 547 engineering students in a public Mexican university. Employing the Winsteps© 4.5.5 (Linacre, 2020), a comparison of the students' scores is made with responses first graded dichotomously and then graded as partial credit. The results show that although scores are not equivalent, these maintain the same order indicated by a correlation near 1. It is concluded that the Rasch & Masters PCM operates efficiently and can be employed even on items with elements not necessarily are concatenated as a sequence of incremental steps, as well as, in areas other than academic performance.

CÓMO CITAR:

Backhoff, E., González-Montesinos, M., Pérez-Garibay, Y. y Ferreyra M. F. (2022). Uso del modelo de crédito parcial de Rasch y Masters en la evaluación de competencias matemáticas. *REICE. Revista Iberoamericana sobre Calidad, Eficacia y Cambio en Educación, 20*(1), 41-55. https://doi.org/10.15366/reice2022.20.1.003

*Contacto: ebackhoff@metrica.edu.com
ISSN: 1696-4713
revistas.uam.es/reice

Recibido: 15 de octubre 2020
1ª Evaluación: 13 de enero 2021
2ª Evaluación: 11 de abril 2021
Aceptado: 27 de junio 2021

1. Revisión de la lieratura

Los avances de la psicometría, las ciencias cognitivas y las tecnologías digitales han permitido que se desarrolle lo que hoy se conoce como evaluación asistida por computadora (EAC), con la que se ha podido superar muchas de las limitaciones que imponen el formato de lápiz y papel y de opción múltiple en las evaluaciones a gran escala. La EAC hace uso de las ventajas que ofrece la tecnología digital, lo que posibilita: 1) formular preguntas que exigen respuestas abiertas a los estudiantes como (p. ej., escribir una ecuación algebraica o balancear una ecuación química), 2) generar una cantidad considerable de preguntas semejantes y, por lo tanto, de exámenes, con el uso de Generadores Automáticos de Ítems (GAI), 3) administrar los exámenes de forma adaptativa de acuerdo con las respuestas (correctas o erróneas) que va emitiendo cada estudiante y 4) calificar las respuestas de los estudiantes de manera parcial, es decir, de acuerdo con el grado en que el alumno responde correctamente los distintos elementos[1] de un reactivo.

Aprovechando las ventajas que ofrece la EAC, en 2020 se terminó el desarrollo del Examen de Competencias Básicas de Matemáticas (Excoba/Matemáticas), cuyo propósito es evaluar los aprendizajes que logran adquirir los estudiantes al término de la educación básica y de la educación media superior (12 grados), con el fin de diagnosticar el nivel de dominio matemático de los estudiantes que ingresan a las instituciones de educación y superior. Con el fin de conocer sus propiedades psicométricas, este instrumento se piloteó con estudiantes de los primeros semestres de diversas carreras de ingeniería de la Universidad Autónoma de Ciudad Juárez (México) y sus resultados se analizaron con el programa WinSteps©, utilizando el Modelo de Crédito Parcial (MCP) de Masters (1982), que es una extensión del modelo original de Rasch (1960) (de aquí en adelante, modelo de crédito parcial de Rasch y Masters). Para conocer el comportamiento del MCP, sus resultados se contrastaron con los resultados que se obtuvieron cuando se utilizó un método de calificación en el que cada uno de los elementos que componen un reactivo se considera como ítem independiente. En síntesis, los resultados del Excoba/Matemáticas se analizaron bajo dos esquemas: 1) considerando los elementos de un reactivo de manera independiente y 2) considerando los elementos de un reactivo de manera integrada (MCP).

El objetivo de este trabajo es doble. Primero, contribuir al campo de la evaluación de las competencias matemáticas, investigando las ventajas y limitaciones que tiene el uso del MCP de Rasch y Masters para calificar las respuestas de los estudiantes en el Excoba/Matemáticas y estimar su nivel de competencia[2]. Segundo, mostrar a las personas interesadas la manera de llevar a cabo el procedimiento de crédito parcial con el programa WinSteps, con instrumentos de respuesta construida o auténtica.[3]

2. Modelo de crédito parcial de Rasch y Masters

Las principales contribuciones a la evaluación del aprendizaje de la aplicación de la Teoría de respuestas al Ítem (TRI) se refieren a la posibilidad de su uso para medir el nivel de habilidad que tiene una persona con relación a cada elemento o característica de la competencia que se desea medir (Nakano y Primi, 2014). La TRI, además, permite analizar de manera detallada la estructura interna de los ítems que se utilizan para evaluar un área de competencia (como es el caso de las matemáticas), indicando cuáles de ellos exigen una menor o mayor capacidad por parte del estudiante. Particularmente, el análisis de Rasch (1960) se ha convertido en uno de los principales recursos técnicos para determinar las propiedades métricas de los ítems de una prueba de logro educativo. Las ventajas de esta metodología han quedado ampliamente establecidas en muchos documentos; véase, por ejemplo a: Wright y Stone (1979), Wright y Masters (1982), Embretson y Reise (2000) y Boone y otros (2014).

[1] Cada uno de los estímulos a los que se debe de responder y que, en conjunto, forman parte integral del ítem. En la literatura especializada, también se les conoce como pasos (*steps*).

[2] Véase, por ejemplo, Henninger (2021), que explora los umbrales de los puntos de corte de las respuestas parciales.

[3] Véase, por ejemplo, Arteaga-Martínez y otros (2018), que analiza la prueba de TIMSS, donde se utilizan respuestas construidas.

El llamado modelo mono-paramétrico de Rasch y sus extensiones derivan su potencial analítico de la aproximación probabilística original. Su formulación axiomática postula que es posible obtener medidas lineales de nivel de rasgo cognitivo, representadas por los parámetros β o habilidades de las personas, de forma independiente de las demandas cognitivas representadas por los parámetros δ o dificultades de los ítems que responden en una prueba de logro académico. Esta propiedad de separabilidad de parámetros es consecuencia directa del hecho que los encuentros entre personas e ítems son eventos probabilísticamente independientes.

El modelo permite establecer una relación entre los diferentes niveles de habilidad de las personas (beta) y las dificultades de los ítems (delta), relación que hace factible identificar diversos patrones de respuesta esperados para cada nivel de beta, mediante el análisis de mapas de ítems (Primi, 2004; Van der Linden y Hambleton, 1997); lo que se ha considerado como una forma más de estudiar la validez de constructo de una prueba. Este procedimiento es la esencia de lo que recientemente se ha denominado el significado de referencia de un ítem (Embretson, 2006; Embretson y Reise, 2000).

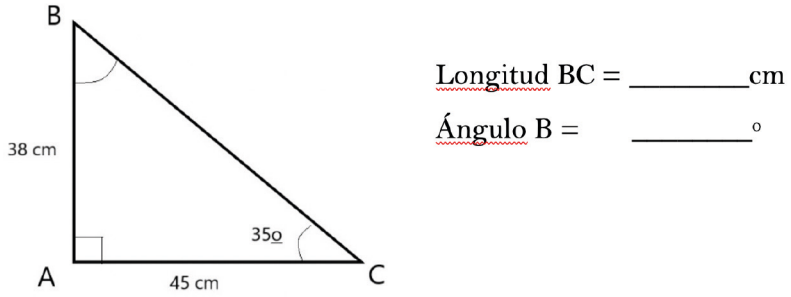
Dado que la TRI establece una relación entre la puntuación en una prueba (o escala) y el puntaje de cada uno de sus ítems, es posible calcular la probabilidad que tendría una persona de responder correctamente un reativo con una dificultad conocida (delta), dado su nivel de competencia (beta). El análisis cualitativo del contenido de cada ítem permite tener una definición más precisa de lo que evalúa la prueba en su conjunto. Asimismo, el análisis jerárquico de la dificultad de los ítems permite inferir cómo se estructura la competencia evaluada con base en sus elementos, con lo que se puede estudiar la validez de constructo de una prueba (Embretson, 2006; Linacre, 1997).

Los ítems de crédito parcial, como algunos que integran el Excoba/Matemáticas, requieren que los estudiantes emitan más de una respuesta para responder un reactivo, de acuerdo con el número de elementos (o pasos) que conforman cada ítem. La Figura 1 muestra el ejemplo de un reactivo con dos elementos o respuestas.

Figura 1

Ejemplo de un reactivo con dos respuestas (o elementos)

Utilizando las propiedades de los triángulos, calcula la longitud del lado BC y la medida del ángulo B



En este caso el modelo formula las expectativas probabilísticas de respuestas correctas para cada una de las partes o elementos de los ítems. La probabilidad de respuesta correcta del sujeto *s* en un ítem *i* y el paso (o elemento) *x* se define por:

$$\Phi_{six} = \frac{\pi_{six}}{\pi_{six-1} + \pi_{six}} = \frac{exp(\beta_s - \delta_{ix})}{1 + exp(\beta_s - \delta_{ix})}$$

$$x = 1,2,\ldots,m_i$$

Esta ecuación es la extensión del modelo Rasch original propuesta por Masters (1982). Concretamente, la diferencia (la habilidad del sujeto menos la dificultad del elemento) contiene la información necesaria para modelar la probabilidad de respuesta correcta del sujeto a cada paso de crédito parcial.

Para un ítem compuesto por *k* pasos se modela la probabilidad de umbral entre pasos (*step thresholds*) por medio de:

$$\pi_{six} = \frac{\exp \sum_{j=0}^{x} (\beta_s - \delta_{ij})}{\sum_{k=0}^{m} \exp \sum_{j=0}^{k} (\beta_s - \delta_{ij})}$$

Esencialmente el MCP trata los pasos de umbral como dicotomías integradas a la estructura del ítem. Esta conceptualización es crucial ya que, como se describe a continuación, el procedimiento computacional requiere reproducir la estructura de cada ítem de crédito parcial en un método de dos fases. La primera se desarrolla para examinar los comportamientos de todos los segmentos de la estructura interna de los ítems en un instrumento frente a las expectativas modeladas. Una vez verificados los índices de bondad de ajuste de los ítems y sus respectivas estructuras de crédito parcial, se retienen aquellos ítems cuyas estructuras satisfacen las expectativas modeladas. La segunda fase se desarrolla para generar las puntuaciones de logro (habilidad) de los sujetos a partir de las puntuaciones sumadas en sus respectivos vectores de respuesta. Las sumatorias de las respuestas correctas de cada sujeto se transforman en unidades logarítmicas de los cocientes de aciertos sobre errores, que son las medidas lineales características del análisis de Rasch.

La modelación de estructuras de crédito parcial en su fase 1 establece las expectativas de respuesta correcta para cada ítem y los segmentos (pasos) internos que lo conforman. Las expectativas probabilísticas se verifican mediante la determinación de los índices de bondad de ajuste para cada segmento interno y para cada ítem en particular. Al igual que en el caso de dicotomías convencionales en ítems binarios, los índices de bondad de ajuste deben mantenerse en el intervalo de 0.80 a 1.30, tanto para el ajuste próximo[4] (INFIT) como para el ajuste externo[5] (OUTFIT). También, según los criterios convencionales cuando los ítems y sus segmentos de estructura interna se mantienen dentro de ese intervalo, los ítems y sus elementos internos son respondidos correctamente solo por los sustentantes que se ubican en el nivel de rasgo requerido por el ítem o por el segmento de estructura bajo análisis. Es decir, en el ajuste interno los pasos de umbral correctos al interior de cada ítem son logrados únicamente por los sustentantes que, de acuerdo con su medida de habilidad, están al nivel requerido por la dificultad de cada paso de umbral. A la inversa, en el ajuste externo los pasos de umbral correctos no son logrados por los sustentantes que no están en el nivel requerido, de acuerdo con la dificultad del cada paso o elemento.

La fase 1 resulta en una valoración de los indicadores psicométricos de los componentes del instrumento, ítem por ítem y elemento por elemento, para determinar su ajuste a las expectativas probabilísticas del modelo de crédito parcial. En la fase 2, los patrones de respuesta de los sustentantes se analizan en forma de puntuaciones sumadas según el número de elementos que cada examinado conteste correctamente. Para este propósito solo se retienen los ítems y elementos de crédito parcial que resultaron estar dentro de los límites establecidos en cuanto a la bondad de ajuste (descritos anteriormente). El resultado de esta fase consiste en las puntuaciones finales de los sustentantes. Estas puntaciones representan con suficiencia el nivel de las competencias a evaluarse y, a partir de ellas, es posible determinar puntos de corte y niveles de logro.

De manera esquemática, en el procedimiento Rasch-Masters para ítems de crédito parcial se establecen expectativas probabilísticas para cada ítem según su estructura interna. Ejemplificando esto con un ítem de cuatro elementos que solicita igual número de respuestas, la estructura interna que se analiza sería la siguiente:

Elementos	0	0,25	0,50	0,75	1
	|———	——|———	——|———	——|——-	|
Umbrales de paso	j_i	k_i	l_i	m_i	

[4] También conocido como ajuste interno

[5] También conocido como ajuste externo

El avance con acierto de cada persona en los cuatro umbrales está definido por la sumatoria de las diferencias entre nivel de habilidad del sustentante β_s y la dificultad de cada uno de los pasos δ_{ij} que es: $\sum_{j=0}^{mi}(\beta_s - \delta_{ij})$

Para cada umbral de paso existe una probabilidad de avance con acierto:

$$p_{i0}(\theta) = 1.0 - p_{i1}^{*}(\theta)$$
$$p_{i1}(\theta) = p_{i1}^{*}(\theta) - p_{i2}^{*}$$
$$p_{i1}(\theta) = p_{i2}^{*}(\theta) - p_{i3}^{*}$$
$$p_{i1}(\theta) = p_{i3}^{*}(\theta) - p_{i4}^{*}$$
$$p_{i4}(\theta) = p_{i4}^{*}(\theta) - 0$$

Estas probabilidades de respuesta correctas se modelan de manera iterada para cada ítem de acuerdo con su estructura, manteniendo la premisa básica del modelo de Rasch: a mayor nivel de habilidad de un estudiante, mayor será su probabilidad de responder correctamente cada uno de los elementos o pasos que conforman un reactivo.

3. Excoba/Matemáticas

Como ya se mencionó, el propósito del Excoba/Matemáticas es evaluar las competencias de matemáticas de los estudiantes que terminan la educación básica y la educación media superior. Los contenidos de este examen están alineados a los aprendizajes esperados que se señalan en los planes y programas de estudio mexicanos, desde 5º grado de primaria hasta 2º grado de bachillerato (Marco Curricular Común[6]). En el Cuadro 1 se muestra la estructura del Excoba/Matemáticas, que se compone de dos grandes bloques, siete ejes temáticos y 184 reactivos (Métrica Educativa, 2020). Los estudiantes que terminan la educación básica (9 grados escolares) responden los 88 reactivos de este nivel educativo, mientras que los alumnos que terminan la educación media superior (12 grados escolares) responden la totalidad de la prueba, es decir, 184 reactivos.

Cuadro 1
Estructura del Excoba/Matemáticas

Nivel educativo	Ejes temáticos	Ítems
Educación Básica	Número, álgebra y variación	51
	Forma espacio y medida	24
	Análisis de datos	13
Educación Media Superior	Pensamiento numérico y lenguaje algebraico	27
	Pensamiento geométrico y trigonométrico	23
	Pensamiento geométrico analítico	18
	Lenguaje variacional: funciones y derivadas	28
Total		184

Adicionalmente, este examen se administra a través de una interfaz computarizada; utiliza reactivos de respuesta construida, semi construida y de selección, en los que los reactivos individualmente le pueden solicitar al estudiante más de una respuesta. Para responder las preguntas abiertas (o de respuesta construida), los alumnos deben de escribir las respuestas en los espacios

[6] En México existe una gran diversidad de planes y programas de estudio del nivel de educación media superior (grados 10 a 12). Todos ellos comparten algunos contenidos comunes, que se imparten en los dos primeros grados.

definidos, como se muestra en la Figura 2 en la que se solicita simplificar una expresión algebraica. La forma de hacerlo es la misma que se utiliza en las calculadoras científicas.

Figura 2
Pantalla del Excoba/Matemáticas que muestra un reactivo de respuesta construida

Por otro lado, algunos reactivos solicitan al estudiante emitir más de una respuesta, en donde se utiliza el modelo de crédito parcial. En este caso, el modelo otorga un valor máximo de un punto por cada reactivo, independientemente, del número de elementos o respuestas solicitadas. Así, un reactivo que solicite una sola respuesta (como el que se ejemplifica en la Figura 2) tendrá un valor de 1 si es correcta o de 0 si es incorrecta; los reactivos que solicitan dos respuestas tendrán un valor de 0,5 puntos cada una; si solicita tres respuestas, cada una tendrá un valor de 0,333 puntos; si pide cuatro, el valor será de 0,25 y así sucesivamente.

Los resultados de los estudiantes se proporcionan en una escala estandarizada, cuyo rango es de 200 a 800 puntos, con una media de 500 unidades y una desviación estándar de 100. Con esta calificación estandarizada, se ubica al estudiante en una categoría de su nivel de logro, la que describe de manera genérica las competencias matemáticas que domina y no domina.

4. Estudio piloto del Excoba/Matemáticas

Para conocer las propiedades métricas de los reactivos del examen y poder establecer los niveles de logro de este instrumento, se realizó un estudio piloto en enero de 2020 con estudiantes voluntarios de los primeros semestres de distintas carreras de ingeniería de la Universidad Autónoma de Ciudad Juárez. Con el fin de lograr que los estudiantes respondieran en las mejores condiciones, evitando cansancio y agotamiento, se conformaron tres versiones de 81 o 82 reactivos cada una, con 30 ítems comunes (o ancla) y 51 o 52 reactivos únicos. Las tres versiones se seleccionaron con el criterio de que cada una abarcara los mismos niveles educativos, ejes temáticos y aprendizajes

esperados, y que tuvieran un nivel de dificultad equivalente (estimado de manera teórica por un grupo de expertos).

En total, se presentaron 547 alumnos a los cuales se les asignó, por turnos, una computadora y una de las tres versiones del examen. Desgraciadamente, el estudio piloto adoleció de varios problemas que limitan su utilidad. Por un lado, los estudiantes no tuvieron la motivación necesaria para responder la totalidad de los reactivos administrados, lo que ocasionó que dejaran muchas preguntas sin responder (valores perdidos). Por otro lado, por razones de logística, las tres versiones del examen no se distribuyeron homogéneamente; así, mientras que la versión A la respondieron 279 estudiantes, las versiones B y C las respondieron solo 142 y 144, respectivamente.

5. Método de calificación y análisis

Los resultados del estudio piloto se analizaron con el paquete estadístico Winsteps© (ver. 4.5.5), utilizando dos procedimientos distintos que, en teoría, deberían dar resultados equivalentes. El primero consistió en analizar los elementos (o pasos) que componen cada reactivo de manera independiente. Es decir, si un ítem requiere dos respuestas del estudiante (como se muestra en la figura 1), cada respuesta se considera de manera aislada, como si fueran dos reactivos distintos. Este método es útil para conocer las propiedades métricas de todos los elementos que componen el examen y, con ello, identificar aquellos que se comportan de manera anómala a fin de poderlos corregir.

Para calificar a los estudiantes, con el primer procedimiento, cada elemento de un reactivo se calificó dicotómicamente (1, 0) para después ponderar su peso de acuerdo con el número de elementos que integran el reactivo. Así, la suma de dos elementos que conforman un ítem fue igual a 1.0 y cada uno de ellos tuvo un valor ponderado de 0.5 puntos. El segundo procedimiento consistió en analizar las respuestas de los reactivos de manera integrada con el MCP. En este caso la ponderación de cada elemento que conforma un reactivo se considera, desde el principio, como parte de la calificación parcial del ítem.

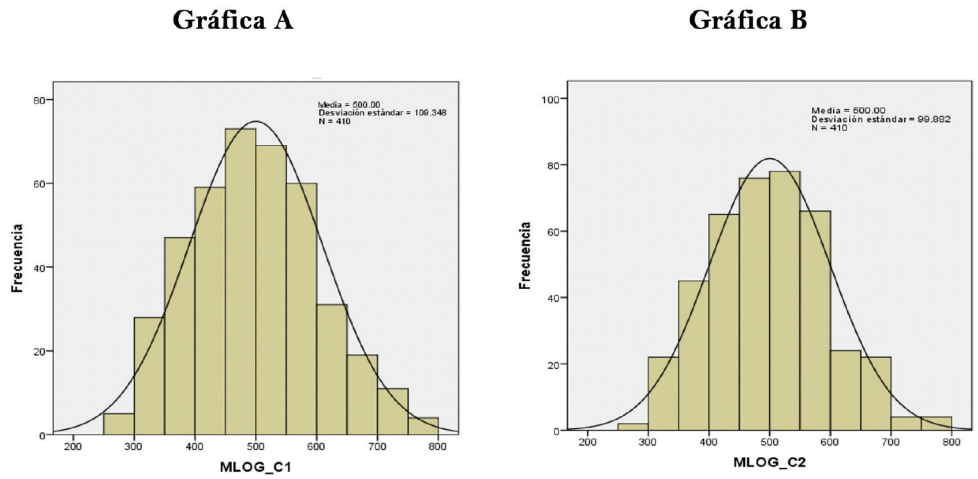
Por otro lado, debido a la gran proporción de respuestas omitidas por los estudiantes, en este trabajo se presentan solo los resultados de los 30 reactivos ancla y los 47 elementos que los componen. Asimismo, del total de la población estudiada (547), se seleccionó una submuestra de 410 alumnos que cumplió con el requisito de haber respondido (o intentado responder) al menos el 50% de los reactivos ancla. La razón de esta decisión, aunque es arbitraria, obedece a tratar de garantizar que los estudiantes hubieran contestado un mínimo de preguntas anclas (al menos, 15 de 30), que son las que se utilizan para que las tres versiones del examen se puedan escalar en una misma métrica. Subir este criterio a más de 50% implicaría perder alumnos en el análisis; mientras que bajarlo significaría tener una métrica menos robusta.

En los anexos 1 y 2 se muestran las instrucciones en WinSteps que se utilizaron para correr los procedimientos antes descritos con los reactivos y estudiantes del estudio.

6. Resultados

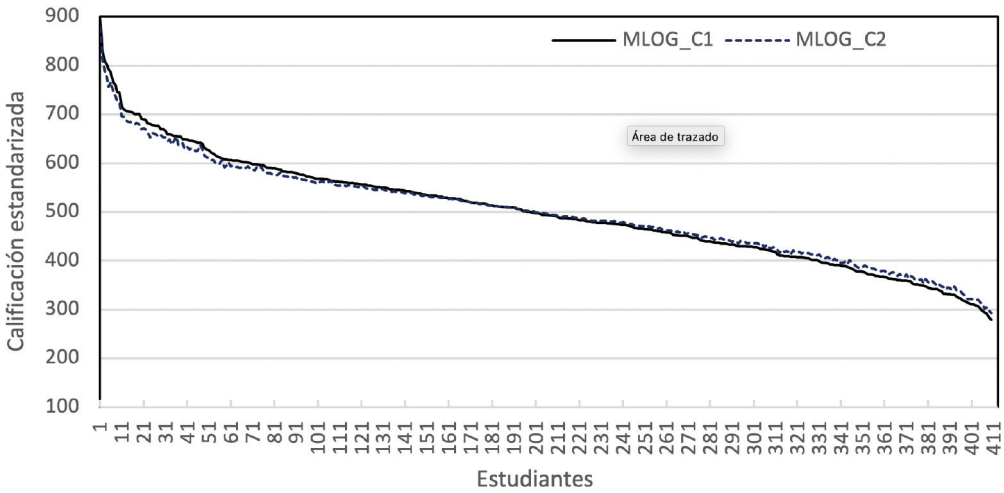
Las Gráficas a y b de la Figura 3 muestran las frecuencias de calificaciones cuando se generan con el procedimiento de los 47 elementos separados (de aquí en adelante, C1) y de los 30 reactivos integrados (de aquí en adelante, C2). Podemos observar que en ambos casos se presenta una distribución de puntuaciones muy parecida, que se asemejan a una curva normal, con una media de 500 puntos, pero con desviaciones estándar ligeramente diferentes: 109.3 (para C1) y 99.8 (para C2).

Figura 3
Distribución de respuestas correctas de los estudiantes cuando se calculan con los 47 elementos separados (a) y los 30 reactivos integrados (b)



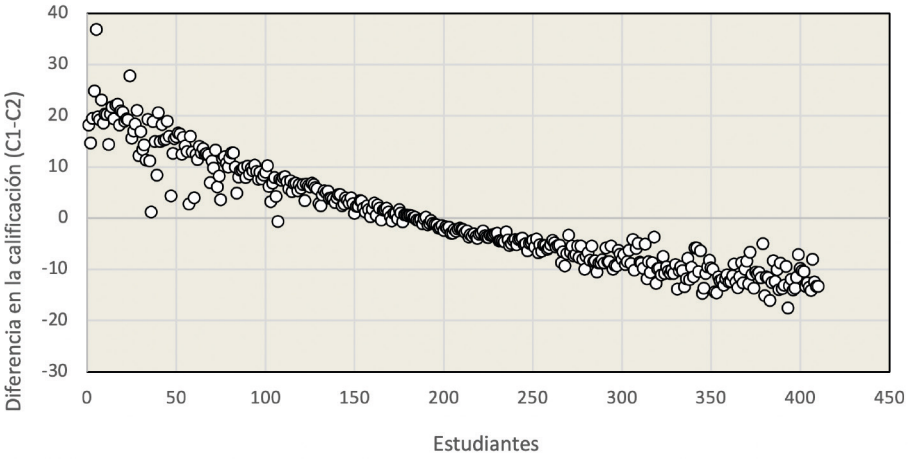
La Figura 4 muestra una comparación de las calificaciones de los 410 estudiantes generadas con ambos métodos (C1 y C2), ordenadas de mayor a menor puntuación. Se podrá observar que estas puntuaciones son muy similares para cada estudiante, con mayores variaciones en los estudiantes que se encuentran en los extremos de la escala de calificación.

Figura 4
Calificaciones de los estudiantes utilizando los métodos C1 (elementos) y C2 (reactivos)



La Figura 5 muestra estas diferencias para cada uno de los 410 estudiantes. Como se puede observar, su inmensa mayoría se encuentra en un rango de ±20 puntos y las diferencias mayores ocurren en los extremos de la escala de habilidad matemática. Sin embargo, estas diferencias no son importantes para calificar a los estudiantes, pues ambos métodos conservan el mismo ordenamiento de alumnos, lo que se comprobó calculando la correlación entre ambas puntuaciones, que fue de 0.999 y estadísticamente significativa (0,001).

Figura 5
Diferencias en las calificaciones generadas por los métodos C1 (elementos) y C2 (reactivos)

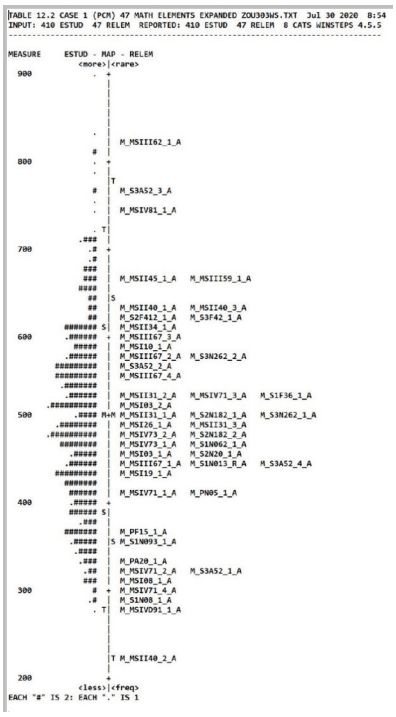


Una vez comprobada la equivalencia de las puntuaciones, se procedió a comparar la distribución de las habilidades de los 410 estudiantes, respecto a las dificultades de los 47 elementos y 30 reactivos. En la figura 6 se muestran los Mapas de Wright para ambos casos: la gráfica A para el caso del análisis por elemento y la B para el análisis por reactivo (MCP). Nótese que con ambos procedimientos los valores medios (M) de las habilidades de los estudiantes (identificados con el carácter #) coincide con los valores medios (M) de las dificultades de los elementos e ítems (identificados con su clave correspondiente) y que en ambos casos las desviaciones estándar de los elementos y reactivos son ligeramente más grandes que las de los estudiantes.

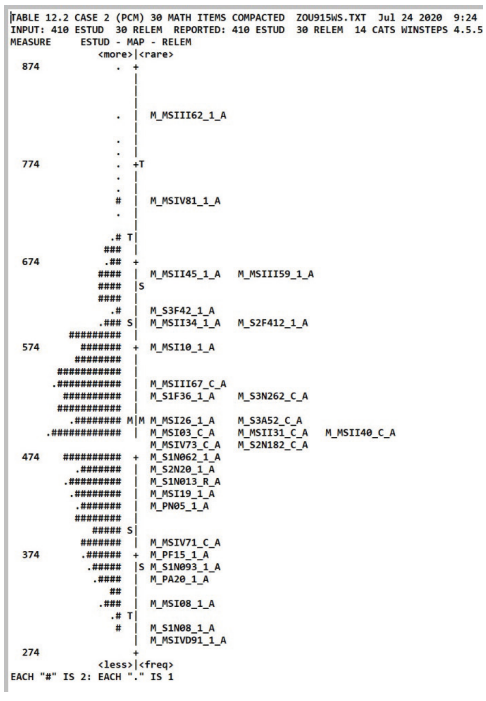
La forma de estas distribuciones de elementos y reactivos es evidencia de que los elementos ancla en su forma desagregada (dicotomías binarias) y en su forma compactada (MCP) están alineados con la distribución de los niveles de habilidad de los 410 estudiantes en la muestra piloto. Esto quiere decir que la gama de dificultad de los reactivos del Excoba/Matemáticas cubre perfectamente la gama las habilidades de los 410 estudiantes.

Figura 6
Comparación de los Mapas de Wright de alineación de 410 sustentantes y 47 elementos (gráfica izquierda) y 30 reactivos (gráfica derecha) ancla Excoba/Matemáticas

Gráfica A



Gráfica B



Finalmente, comparamos los dos métodos de análisis para identificar el comportamiento anómalo de los elementos y reactivos del Excoba/Matemáticas; es decir, aquellos reactivos que no se ajustan a los parámetros esperados y que requieren revisarse y mejorarse. Específicamente, se buscó que los intervalos de ajuste interno (*infit*) y externo (*outfit*) no fueran menores a 0.8 ni mayores a 1.30, siempre que su indicador estandarizado (Sign Z) sea igual o mayor a 2.0. El Cuadro 2 (para elementos) y 3 (para reactivos) muestran los resultados de este análisis, solo para aquellos casos en los que se identificó algún problema de desajuste. En el Cuadro 2 se aprecia que ocho elementos cumplen con este criterio de desajuste. Estos elementos son: 2 y 3 del ítem 67 (M_MSIII67_2_A, M_MSIII67_3_A); 2 y 3 del ítem 71 (M_MSIV71_2_A, M_MSIV71_3_A); 1 del ítem 12 (M_MS2NI82_1_A) y 1, 2 y 4 del ítem 52 (M_S3A52_1_A, M_S3A52_2_A y M_S3A52_4_A).

Cuadro 2
Indicadores psicométricos de los elementos de los reactivos del Excoba/Matemáticas que presentaron algún desajuste

Elemento	Peso	Delta	Err. Est.	Ajuste Interno (AI)	AI Sig. Z	Ajuste externo (AE)	AE Sig. Z
M_MSIII67_2_A	0,25	573,62	13,81	1,3819	5,4614	1,6443	5,6316
M_MSIII67_3_A	0,25	595,15	14,20	1,2939	3,9813	1,4996	4,0415
M_MSIV71_2_A	0,25	325,32	15,45	1,112	1,3511	1,6677	3,1117
M_MSIV71_3_A	0,25	528,02	13,17	1,2245	3,8812	1,3397	3,8613
M_S2F412_1_A	1,00	612,94	12,44	1,249	3,9012	1,5025	4,5315
M_S3A52_1_A	0,25	326,6	14,03	1,147	1,8711	1,4301	2,3814
M_S3A52_2_A	0,25	558,62	11,92	1,2376	4,4312	1,4322	5,0514
M_S3A52_4_A	0,25	448,46	11,74	1,3435	6,6213	1,5808	6,1216

Por su parte, en el Cuadro 3 se muestran los cinco reactivos que en los análisis se salieron del rango esperado: M_MSIII31_C_A, M_MSIII67_C_A, M_MSIV71_C_A, M_S2F412_1_A y M_S3A52_C_A.

Cuadro 3
Indicadores psicométricos de los reactivos del Excoba/Matemáticas que presentaron algún desajuste

Ítem	Peso	Delta	Ajuste Interno (AI)	AI Sig. Z	Ajuste Externo (AE)	AE Sig. Z
M_MSII31_C_A	0,33	493,37	1,7668	6,6618	1,8952	4,9019
M_MSIII67_C_A	0,25	534,75	1,9336	9,3019	2,0258	9,882
M_MSIV71_C_A	0,25	388,62	1,7292	7,6917	1,9098	8,3519
M_S2F412_1_A	1	608,15	1,1968	3,2312	1,3845	3,8614
M_S3A52_C_A	0,25	505,05	1,3486	4,5913	1,5113	6,2915

Comparando los resultados de ambos métodos, encontramos que coinciden en que los ítems 67, 71, 12 y 52 presentan algún grado de desajuste. La excepción fue el ítem 31, cuyo desajuste solo se identificó con el procedimiento MCP, lo que podría sugerir que sus elementos deben ser analizados como ítems individuales, por algún posible problema de dependencia.

7. Discusión

El propósito de este trabajo fue doble. Por un lado, describir el modelo de crédito parcial (MCP) propuesto por Masters (1982), que es una extensión del modelo original de Rasch (1960), utilizando el programa WinSteps©

4.5.5 (Linacre, 2020). Por otro lado, mostrar los resultados que genera dicho modelo con un caso real. El estudio consistió en comparar los resultados que genera el método MCP (calificación de las respuestas parciales de cada reactivo) con el método simple de considerar las respuestas a un ítem como reactivos independientes (o dicotómicos). Con el programa WinSteps, se realizaron dos tipos de análisis, asumiendo que las respuestas parciales a un reactivo: 1) son dicotomías independientes que se califican individualmente (método C1) y 2) son puntuaciones parciales que se califican conjuntamente (método C2). Para realizar estos análisis, se utilizó la información proveniente de un estudio piloto del Examen de Competencias Básicas de Matemáticas (Excoba/Matemáticas) (Métrica Educativa, 2020), realizado en enero de 2020, con estudiantes de distintas carreras de ingeniería de la Universidad Autónoma de Ciudad Juárez.

Básicamente, se realizaron dos tipos de análisis: 1) la estimación de las calificaciones de los estudiantes y 2) la identificación de los reactivos (y elementos) que muestran un desajuste con el modelo teórico. En cuanto a las calificaciones de los estudiantes, se encontró que ambos métodos generan puntuaciones equivalentes que, aunque no idénticas, su ordenamiento es el mismo; razón por lo que la correlación entre ambas es cercana a 1. La diferencia entre las puntuaciones se debe a que la varianza de los elementos es mayor que de los reactivos, debido a que en el primer caso se analizaron 47 respuestas independientes y en el segundo se analizaron 30 reactivos integrados.

En cuanto a las propiedades métricas de los reactivos, se encontró que con ambos métodos se pudieron identificar a aquellos con un desajuste significativo. Con el método C1 se identificaron ocho elementos anómalos, pertenecientes a cuatro reactivos; mientras que con el método C2 se identificaron cinco ítems anómalos: los cuatro identificados con el primer método, más uno adicional. En síntesis, se puede concluir que el MCP genera resultados muy similares al método de calificación individual de reactivos.

Estos resultados son consistentes con los hallazgos de Andrich (2016) donde mediante estudios de simulación, se muestra que el tratar los elementos de la estructura interna de los reactivos como dicotomías se obtienen estimaciones de los parámetros de dificultad de los umbrales de respuesta correcta y de las habilidades de las personas de manera equivalente a los que se obtienen tratando los patrones de respuesta como dicotomías independientes. Esta aproximación al análisis de crédito parcial es justamente la que se ha seguido en este estudio. Por lo anterior, concluimos que en casos como el analizado, la técnica psicométrica más apropiada es la que se aplicó en este estudio, ya que produce la estimación de las calificaciones de los estudiantes, previa identificación exhaustiva de los elementos y reactivos que muestran ajuste o desajuste con el modelo teórico.

Un aspecto importante de mencionar en este estudio es el concepto de calificación parcial que, en teoría, está pensado para ítems que requieren de dos o más pasos para llegar a una solución. Por ejemplo, un problema de física puede solicitarle al estudiante tres respuestas: la fórmula, su aplicación y la unidad de medida correspondiente. Cada paso del reactivo cuenta en su calificación, de tal manera que el reactivo se puede calificar correctamente (si tiene tres aciertos), incorrectamente (si no tiene ningún acierto) o parcialmente (si tiene una o dos respuestas correctas). Ahora bien, en el Excoba/Matemáticas no se utiliza el concepto de pasos, sino el de elementos, cuyas respuestas no están necesariamente secuenciadas o concatenadas. Por ejemplo, en la Figura 1 se le solicita al estudiante que calcule la longitud de un lado del triángulo y la apertura de un ángulo, lo que se puede realizar de manera independiente. Esta particularidad es importante de mencionar, dado que en el caso del Excoba/Matemáticas no existe el efecto de dependencia de los elementos que conforman un reactivo.

Los resultados de este estudio muestran que MCP de Rasch y Masters opera de manera normal en reactivos de crédito parcial cuyos elementos de estructura son en realidad dicotomías independientes, aun cuando los elementos estén conceptualmente relacionados. Este hallazgo nos permite resaltar el potencial analítico del MCP que puede extenderse hacia reactivos cuyas estructuras internas no están concebidas como pasos ordenados en secuencias incrementales.

Una de las limitaciones de este trabajo es que se basó en los resultados de un estudio piloto del Excoba/Matemáticas, donde participaron estudiantes que no fueron adecuadamente motivados para responder la totalidad de la prueba. Una segunda limitación es que del total de reactivos que contiene este examen (184), se utilizaron solamente los 30 ítems reactivos ancla y que, en conjunto, solicitaron al estudiante emitir 47 respuestas. De los 547 estudiantes, se seleccionó una submuestra de 410, que cumplieron con el requisito de haber intentado responder al menos el 50% de estos 30 reactivos. No obstante, estas limitaciones, la compa-

ración realizada proporciona suficiente información para poder afirmar que el MCP es un método adecuado para calificar las competencias cognitivas de los estudiantes, así como para identificar reactivos de una prueba que deben revisarse y mejorarse. De cualquier manera, habría que replicar estos resultados en un estudio futuro en el que los estudiantes respondan todas las preguntas del examen, poniendo su mayor esfuerzo. Igualmente, habría que realizar estudios semejantes con otro tipo de competencias escolares, como podría ser la comprensión de lectura (Primi, 2004) u otra.

Finalmente, es importante señalar que el uso del MCP no se ha limitado a evaluaciones de competencias escolares, sino que también se ha utilizado con instrumentos para evaluar competencias intelectuales (Freitas et al., 2015; Primi, 2004), habilidades musicales (Wesolowski et al., 2016) y de otros tipos. En esta dirección se encuentra un área de oportunidad para realizar investigaciones complementarias a los análisis realizados de logro académico, como los que se reportan en este trabajo.

Referencias

Andrich, D. (2016). Inference of independent dichotomous responses in the polytomous rasch model. *Rasch Measurement Transactions*, *30*(1), 1566-1569.

Arteaga Martínez, B., Navarro Asencio, E., Fraile Rey, A. y Ramos Alonso, P. (2018). Adaptación de la prueba TIMSS para la evaluación de la competencia matemática en alumnos de magisterio. *Bordón. Revista de Pedagogía*, *70*(3), 95-113.

Boone, W. J, Staver, J. R. y Yale, M. R. (2014). *Rasch analysis in the human sciences.* Springer. https://doi.org/10.1007/978-94-007-6857-4

Embretson, S. E. (2006). The continued search for nonarbitrary metrics in psychology. *American Psychologist*, *61*(1), 50-55. https://doi.org/10.1037/0003-066X.61.1.50

Embretson, S. E. y Reise, S. P. (2000). *Item response theory for psychologists*. Lawrence Erlbaum.

Freitas, S., Prieto, G., Simões, M. R. y Santana, I. (2015). Scaling cognitive domains of the montreal cognitive assessment: An analysis using the partial credit model. *Archives of Clinical Neuropsychology*, *30*(5), 435-447. https://doi.org/10.1093/arclin/acv027

Henninger, M. (2021). A novel partial credit extension using varying thresholds to account for response tendencies. *Journal of Educational Measurement, 58*(1), 104-129. http://doi.org/10.1111/jedm.12268

Linacre, J. M. (1997). Kr-20/Cronbach alpha or rasch reliability: Which tells the "truth"? *Rasch Measurement Transactions*, *11*(3), 580-581.

Linacre, J. M. (2020). *Winsteps® (Versión 4.5.5)*. Winsteps.com.

Masters, G. N. (1982). A rasch model for partial credit scoring. *Psychometrika, 47*(2), 149-174. https://doi.org/10.1007/BF02296272

Métrica Educativa, A. C. (2020). *Excoba/Matemáticas/ES*. https://metrica.edu.mx/examenes/examenes-diagnosticos/exumat/excoba-matematicas-es/

Nakano, T. y Primi, R. (2014). Rasch-master's partial credit model in the assessment of children's creativity in drawings. *The Spanish Journal of Psychology, 17*(35), 1-16. http://doi.org/10.1017/sjp.2014.36

Primi, R. (2004). Avanços na interpretação de escalas com a aplicação da teoria de resposta ao item. *Avaliação Psicológica, 3*(1), 53-58.

Rasch, G. (1960). *Probabilistic models for some attainment and intelligence tests.* Danmarks Pædagogiske Institut.

Van der Linden, W. J. y Hambleton, R. K. (1997). *Handbook of modern item response theory*. Springer. http://doi.org/10.1007/978-1-4757-2691-6

Wesolowski, B., Wind, S. y Engelhard, G. (2016). Examining rater precision in music performance assessment: An analysis of rating scale structure using the multifaceted rasch partial credit model. *Music Perception: An Interdisciplinary Journal, 33*(5), 662-678. https://doi.org/10.1525/mp.2016.33.5.662

Wright, B. D. y Masters, G. N. (1982). *Rating scale analysis*. MESA Press.

Wright, B. D. y Stone, M. H. (1979). *Best test design*. MESA Press.

Anexo 1. Archivo de control Winsteps© para análisis de 47 elementos dicotómicos

```
! Inicia la secuencia de instrucciones
&INST
TITLE="CASE 1 (PCM) 47 MATH ELEMENTS EXPANDED "
! Especificación de base de datos
DATA=.\01DTEXPA.DAT
XWIDE=1
PERSON=ESTUD
NAME1=1
NAMELEN=8
NITEMS=47
ITEM=RELEM
ITEM1=9
! Agrupamientos por tipo de reactivo
ISGROUPS=BBAAAACCCACCCAAADDDDDDDDDBBAAAAAAAAAAABBADDDDABB
! Códigos validos en patrones de respuesta
CODES=01
IVALUEA=01
IVALUEB=01
IVALUEC=01
IVALUED=01
¡Los espacios en blanco se tratan como items no presentados
MISSCORE=-1
¡Mostrar resultados incluyendo items extremos
TOTALSCORE=YES
!Emplear la distribución normal teórica para estandarización
LOCAL=NO
! Centrar la escala de items en media de 500 y desviación estándar de 100
UIMEAN=500
USCALE=100
!Esquema de ponderaciones para los elementos y reactivos según su tipo.
IWEIGHT=*
01      0.50
02      0.50
03      1.00
04      1.00
05      1.00
06      1.00
07      0.33
...     (se omiten los pesos de todos los elementos de los ítems para ahorrar espacio)
47      0.50
*
!Especificaciones para información de salida
DISCRIM=YES
ASYMPTOTE=YES
IDELQ=YES
PDELQ=YES
! Fin de la secuencia de instrucciones
&END
¡ Etiquetas de los reactivos (por razón de espacio se omite el listado completo en este ejemplo)
M_MSI03_1_A
M_MSI03_2_A
M_MSI08_1_A
M_MSI10_1_A
.
END NAMES
```

Anexo 2. Archivo de control Winsteps© para análisis de 30 reactivos de crédito parcial

```
&INST
Title="CASE 2 (PCM) 30 MATH ITEMS COMPACTED "
!Especificación de la base de datos
DATA=.\02DTCOMPA.DAT
XWIDE=1
PERSON=ESTUD
NAME1=1
NAMELEN=8
NITEMS=30
ITEM=RELEM
ITEM1=9
!Agrupamientos por tipo de reactivo
ISGROUPS=BAAAACACAAADDBAAAAAAAAAAABADAB
!Códigos validos en patrones de respuesta
CODES=01234
IVALUEA=01
IVALUEB=012
IVALUEC=0123
IVALUED=01234
! Los espacios en blanco se tratan como items no presentados
MISSCORE=-1
!Mostrar resultados incluyendo items extremos
TOTALSCORE=YES
!Emplear la distribución normal teórica para la estandarización
LOCAL=NO
! Centrar la escala de persomas en media de 500 y desviación estándar de 100
UPMEAN=500
USCALE=100
! Esquema de ponderación de los reactivos
IWEIGHT=*
01      0.50
02      1.00
03      1.00
04      1.00
05      1.00
06      0.33
07      1.00
08      0.33
09      1.00
10      1.00
.
.               (se omiten los pesos de todos los elementos de los ítems para ahorrar espacio)
28      0.25
29      1.00
30      0.50
*
! Especificación de la información de salida
DISCRIM=YES
ASYMPTOTE=YES
IDELQ=YES
PDELQ=YES
! Fin de la secuencia de instrucciones
&END
! Etiquetas de los reactivos (por razón de espacio se omite el listado completo en este ejemplo)
M_MSI03_C_A
M_MSI08_1_A
M_MSI10_1_A
M_MSI19_1_A
.
.
END NAMES
```

Breve CV de los/as autores/as

Eduardo Backhoff Escudero

Fue profesor de Psicología en la Universidad Nacional Autónoma de México (1975-1979); investigador y director del Instituto de Investigación y Desarrollo Educativo de la Universidad Autónoma de Baja California (1988-2013); director científico de la Revista Electrónica de Investigación Educativa; director de Pruebas y Medición y consejero presidente de la Junta de Gobierno del Instituto Nacional para la Evaluación de la Educación. Miembro del Sistema Nacional de Investigadores (1990-). Fue representante de México en el proyecto PISA y asesor del proyecto TALIS. Autor de 120 de artículos de investigación, 30 capítulos y 25 libros en materia de evaluación educativa. Su campo de interés es el Especialista en diseño y validación de pruebas asistidas por computadora. Actualmente es presidente de Métrica Educativa, A. C. Email: ebackhoff@metrica.edu.com

ORCID ID: https://orcid.org/0000-0001-7267-4774

Manuel Jorge González Montesinos

Realizó sus estudios de doctorado en el Departamento de Psicología Educativa de la Universidad de Arizona (2000-2004). Fue profesor-investigador en Metodología y Medición de la Universidad de Sonora, México (1979 -2019). Ha sido investigador invitado del Instituto Nacional para la Evaluación de la Educación (2008), consejero técnico del Instituto Nacional para la Evaluación de la Educación (2009-2012) y director general adjunto del INEE en Sonora (2015-2019). Profesor invitado de: Universidad Autónoma de Baja California (2007), Universidad de Valencia (2009, 2011) y Universidad Anáhuac, Doctorado en Evaluación Educativa, (2016 y 2021). Actualmente se desempeña como Director de Posgrado e Investigación del Centro Regional de Formación Profesional Docente de Sonora, México.
Email: manuelgm4@gmail.com

ORCID ID: https://orcid.org/0000-0002-7168-4046

Yadira Pérez Garibay

Realizó estudios de Maestría en Ciencias Educativas en la Universidad Autónoma de Baja California, México (2008-2011). Ha desarrollado instrumentos de evaluación educativa en la Universidad Autónoma de Baja California (2010-2013). Ha sido profesora de Álgebra y Cálculo en la misma universidad (2004-2014); profesora de Matemáticas y Física en el Colegio de Bachilleres del Estado de Baja California, México, y en el Centro de Educación Media Superior a distancia Punta Colonet (2009-2010). Actualmente, se desempeña como investigadora asociada en Métrica Educativa, A.C., México.
Email: yperez@metrica.edu.mx

ORCID ID: https://orcid.org/0000-0001-6675-8699

María Fabiana Ferreyra

Ha sido profesora de Matemáticas en la Escuela Superior de Comercio y Escuela Técnica, Argentina (1988.2001); profesora de Álgebra en el Instituto Nacional Superior de Profesorado Técnico, Argentina, (1992-1996); adscripción a la cátedra Álgebra III en el Instituto Nacional Superior del Profesorado Dr. Joaquín V. González, Argentina, (1989-1990); investigadora en Evaluación educativa y Medición, de la Universidad Autónoma de Baja California, México (2007). Realizó una estancia de Investigación en Matemáticas en la Universidad de Shiga, Japón (1997-1998). Ha desarrollado instrumentos de evaluación educativa en la Universidad Autónoma de Baja California, México, (2009-2013). Actualmente, se desempeña como investigadora asociada en Métrica Educativa A.C., México. Email: fferreyra@metrica.edu.mx

ORCID ID: https://orcid.org/0000-0002-5254-1752