FACULTAD DE INGENIERÍA Y CIENCIAS APLICADAS
UNIVERSIDAD CENTRAL DEL ECUADOR

INGENIO
2023, VOL 6(2), JULIO-DICIEMBRE, PP. 45-54

REVISTA INGENIO

Validación de un Test de Matemática Aplicado a Estudiantes que Ingresan a la Educación Superior, Empleando el modelo de Rasch

Validation of a Mathematics Test Applied to Students Starting Higher Education, Using the Rasch Model

Edgar Valdemar Guamán Tenezaca | iD Instituto Superior Universitario Central Técnico (Ecuador)
Miguel Alonso Murillo Noblecilla | iD Instituto Superior Universitario Central Técnico (Ecuador)
Javier Alexander Castro Haro | iD Instituto Superior Universitario Central Técnico (Ecuador)

ARTICLE HISTORY
Received: 20/01/2023
Accepted: 02/05/2023

PALABRAS CLAVE
Modelo matemático de Rasch, teoría de respuesta al ítem, habilidad, dificultad, distribución estadística.

RESUMEN
El modelo de Rasch aplicado en la calibración de un instrumento de evaluación válido y confiable, el cual consiste en un test de diagnóstico de 20 ítems de matemática, tomado previo a un curso de nivel cero en el Instituto Superior Universitario Central Técnico, entre los períodos 2020-I a 2022-I. Con este test se evaluó a 695 estudiantes en los períodos 2020-I, 2020-II, 2021-I y 2021-II. Posteriormente, se identificaron los ítems del instrumento de evaluación que no son descritos por el modelo de Rasch con una confiabilidad del 65%, mismos que se procedieron a corregir; luego, con el nuevo test corregido, se evaluó a otros 100 estudiantes en el período 2022-I, obteniendo una confiabilidad del test de 90%. A partir de estos resultados se generan las curvas características de dichos ítems y a través de las distribuciones de Pearson y ji-cuadrada se identifica a aquellos que no se ajustan al modelo. Utilizando los parámetros arrojados por el modelo de Rasch se procede a la simulación de las notas y se compara con las reales obtenidas por los estudiantes. Así también, el modelo ha permitido identificar a 133 estudiantes con bajo nivel de habilidad de los cuales 119 corresponden al test original y 14 al test corregido. Para los análisis estadísticos se utilizó el software R.

KEY WORDS
Rasch mathematical model, item response theory, ability, difficulty, statistical distributions.

ABSTRACT
In this test, the Rasch model was applied for the calibration of a valid and reliable assessment instrument. Between the periods 2020-I to 2022-I the diagnostic test consisted of a 20-item mathematics questionnaire taken prior to a level zero course at «Instituto Superior Universitario Central Técnico». With this test, 695 students had been evaluated in the periods 2020-I, 2020-II, 2021-I and 2021-II; subsequently, the items of the evaluation instrument that were not described by the Rasch model were identified with a reliability of 65%, which were corrected. Finally, with the new corrected test, another 100 students had been evaluated in the period 2022-I, obtaining a test reliability of 90%. From these results, the characteristic curves of these items were generated, applying Pearson and chi-square distributions, those that did not fit the model were identified. Using the parameters obtained by the Rasch model, the grades were simulated and compared with the actual grades obtained by the students. Thus, the model has made it possible to identify 133 students with a low level of ability, of which 119 correspond to the original test and 14 to the corrected test. R software was used for the statistical analysis.

I. INTRODUCCIÓN

A partir del año 2019, se implementó en el Instituto Superior Universitario Central Técnico (ISUCT) un curso de nivel cero, que dura aproximadamente dos semanas; esto debido a las deficiencias en matemática que presentan los estudiantes que ingresan a este centro de educación superior [1].

Al inicio de este curso de nivel cero, se ha examinado a los estudiantes (que en adelante se denominarán sustentantes) a través de un test de diagnóstico (original) de matemática, que constan de 20 ítems de opción múltiple. Durante los períodos 2020-I, 2020-II, 2021-I, 2021-II y 2022-I, se han recolectado los resultados de 795 sustentantes que rindieron el test.

Para el análisis cuantitativo de los resultados de este test hay dos enfoques que se puede considerar, uno de ellos es la teoría clásica de los test (TCT), que se origina

https://doi.org/10.29166/ingenio.v6i2.4548

pISSN 2588-0829
eISSN 2697-3243
fing.revista.ingenio@uce.edu.ec

debido a la necesidad de medir las diferencias entre sustentantes, considerando sus atributos o características particulares [2]; es así que la TCT se caracteriza mediante la ecuación:

$$X_0=X_v+e \qquad (1)$$

donde X_0 representa la puntuación observada de un sustentante al aplicarle un test, mientras que X_v indica la puntuación verdadera que consiste en el límite hacia el cual convergería la puntuación observada si se aplicara al sustentante infinitas mediciones. Finalmente, e indica el error de medida, que es la diferencia entre la puntuación observada y la verdadera.

Además, la TCT se fundamenta en tres supuestos básicos establecidos por Spearman en 1904, tal como se menciona en [3], [4]; primero, que la esperanza del error es cero; segundo, la correlación entre el valor observado y verdadero es cero, y tercero, que los errores no se relacionan entre sí.

Sin embargo, el valor observado no es confiable debido a la intervención de factores que perturban la medición, por ejemplo: el instrumento, el sustentante o la situación. Es decir, que algunas propiedades psicométricas de los tests como la dificultad de los ítems o la fiabilidad del test están en función de la muestra de sustentantes utilizada para su validación, como se discute en Mateo y Martínez [5].

Por lo que para solventar las dificultades de la TCT, se ha desarrollado una teoría complementaria, que se conoce como teoría de respuesta al ítem (TRI), fundamentada en los trabajos de Thurstone en 1925, Lawley en 1943, Tucker en 1946 y la síntesis importante que compila los cimientos definitivos de la TRI lo realiza Lord en 1952 y Georg Rasch en 1960; en estos trabajos se formulan modelos mucho más complejos y robustos que el modelo lineal (Eq. 1) de la TCT, tal como se recoge en [6].

El principal objetivo de la TRI, es obtener medidas que sean invariantes respecto a los sustentantes y a los instrumentos de medida o test. Por lo que, para concretar este objetivo, la familia de modelos de la TRI reconoce que la probabilidad de un sustentante en acertar o no un ítem, queda determinada en función de la posición de dicho sustentante en el rasgo latente o habilidad del mismo (notado por θ), y por uno o más atributos del ítem, como puede ser: la dificultad β, la discriminación , el azar, entre otros.

Uno de los modelos de la TRI de mayor aplicación, es el denominado modelo de Rasch, que posee dos supuestos básicos, la unidimensionalidad del espacio latente y la independencia local[3], [5], [7]-[9].

Es así que se considera la variable aleatoria X_{ik} que toma el valor 1 en caso de acertar el sustentante el ítem i, caso contrario toma el valor 0. Luego, la heterogeneidad de la información se expresa como la probabilidad de que esta variable aleatoria tome los valores 0 o 1. Si se representa por π la probabilidad de que la persona k responda afirmativamente el ítem i, esto es $P(X_{ik}=1)=\pi$; caso contrario $P(X_{ik}=0)=1-\pi$. Por consiguiente, se dice que la variable aleatoria X_{ik} sigue una distribución de Bernoulli de parámetro π, que se nota $X_{ik} \sim B(\pi)$.

Por otro lado, se define Ω_{ik} como el cociente entre la probabilidad que el sustentante k acierte el ítem i, respecto a la probabilidad de que lo responda de forma incorrecta, es decir,

$$\Omega_{ik}=\frac{P(X_{ik}=1)}{P(X_{ik}=0)} \Leftrightarrow \Omega_{ik}=\frac{\pi_{ik}}{1-\pi_{ik}} \qquad (2)$$

En 1960 Georg Rasch, en su afán de separar la peculiaridad del sustentante y la peculiaridad del ítem, propone que

$$\Omega_{ik}=\frac{\delta_k}{\tau_i} \qquad (3)$$

donde δ_k es la característica del sustentante k, mientras que τ_i es la característica del ítem i.

En (Eq. 3), al fijar τ_i implica que $\Omega_{ik}>1$ o $\Omega_{ik}<1$ para δ_k grandes o pequeños, respectivamente, por lo que se considera a δ_k como la habilidad del sustentante k respecto al valor fijo τ_i. Así, también al fijar δ_k los valores de Ω_{ik} pueden ser mayores o menores a 1 dependiendo de si τ_i toma valores pequeños o grandes, respectivamente, por lo que se considera a τ_i como la dificultad del ítem i respecto al valor fijo δ_k.

Sin embargo, al considerar un valor fijo de Ω_{ik}, se observa que no existen valores únicos tanto de δ_k como de τ_i correspondientes para dicho valor, esto indica que la habilidad y dificultad no están bien definidos. Por lo que en [10], se define un ítem estándar 1 con la siguiente restricción,

$$\tau_i = 1 \qquad (4)$$

de donde (Eq. 3) se reduce a:

$$\frac{P(X_{ik}=1)}{P(X_{ik}=0)}=\Omega_{ik}=\delta_k, \qquad \forall k=1,\ldots,m \qquad (5)$$

Esto significa que la habilidad del sustentante k se define como el cociente de la probabilidad que este sustentante acierte el ítem estándar respecto a la probabilidad de que lo falle. Así, la dificultad del ítem i corresponde al odd

ratio entre el ítem 1 y el ítem *i*, para cada sustentante k, es decir,

$$\frac{\frac{P(X_{1k}=1)}{P(X_{1k}=0)}}{\frac{P(X_{ik}=1)}{P(X_{ik}=0)}} = \frac{\Omega_{1k}}{\Omega_{ik}} = \frac{\delta_k}{\frac{\delta_k}{\tau_i}} = \tau_i \quad (6)$$

De (Eqs. 5 y 6) y se sigue que,

$$\delta_i > \tau_i \Leftrightarrow \frac{P(X_{1k}=1)}{P(X_{ik}=1)} > \frac{\frac{P(X_{1k}=1)}{P(X_{1k}=0)}}{\frac{P(X_{ik}=1)}{P(X_{ik}=0)}} \Leftrightarrow P(X\_ik=1)>P(X\_ik=0) \quad (7)$$

De (Eq. 7) se concluye que el sustentante *k* tiene una habilidad superior a la dificultad del ítem si y solo si su probabilidad de acertar en la respuesta correcta es mayor a la probabilidad de no acertar.

Luego, de (Eqs. 2 y 3) se tiene que,

$$\Omega_{ik} = \frac{\pi_{ik}}{1-\pi_{ik}} \Leftrightarrow \Omega_{ik} - \Omega_{ik}\,\pi_{ik} = \pi_{ik} \Leftrightarrow \pi_{ik} = \frac{\frac{\delta_k}{\tau_i}}{1+\frac{\delta_k}{\tau_i}} \quad (8)$$

por consiguiente, al reparametrizar (Eq. 8) a través de $\delta_k = e^{\theta_k}$ y , se tiene el modelo siguiente,

$$P(X_{ik}=1) = \frac{\frac{e^{\theta_k}}{e^{\beta_i}}}{1+\frac{e^{\theta_k}}{e^{\beta_i}}} \Leftrightarrow P(X_{ik}=1) = \frac{e^{\theta_k - \beta^i}}{1+e^{\theta_k - \beta_i}} \quad (9)$$

donde $i=1,\ldots,n$ y $k=1,\ldots,m$, se considera m el número de sustentantes y n el número de ítems que componen el test. Así, (Eq. 9) representa el modelo de Rasch, que es un modelo de un parámetro [2], [10], cuyo grafo se denomina Curva Característica del Ítem (CCI) que se representa en la figura 1 (ver Figura 1).

Este estudio, tiene como principal objetivo dar solución a dos problemas: el primero, crear un instrumento de evaluación válido y confiable; en segundo lugar, medir la habilidad de los sustentantes e identificar a los sustentantes con altas deficiencias en matemática para facilitar esta información a la coordinación de Bienestar Estudiantil, y que esta instancia a su vez realice un seguimiento adecuado y se eviten posibles deserciones estudiantiles.

Para abordar estos dos problemas, se considera el enfoque establecido en investigaciones anteriores, [8], [11], [12] en donde se aplica el modelo de Rasch para la calibración del test, obteniendo de esta forma la dificultad de los ítems (θ_k) y la habilidad de los sustentantes (βi). Sin embargo, en este trabajo se realiza un análisis adicional (para la construcción de un adecuado instrumento de evaluación) el cual consiste en determinar el nivel de mejora del test original utilizando un test corregido, para esto se procede a la corrección de los ítems anómalos identificados en el primer test, posteriormente con el nuevo test corregido se evalúa a un grupo de 100 sustentantes del período 2022-I del ISUCT, finalmente se compara los resultados de ambos test, tanto del original como del corregido.

2. Método

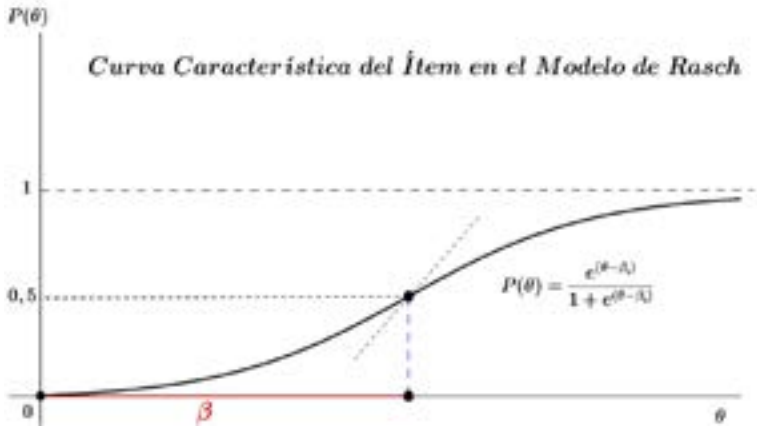
El método considerado en este trabajo es cuantitativo con enfoque descriptivo, esta selección se basa en las respuestas emitidas por los 795 sustentantes a cada uno de los ítems que componen el instrumento de evaluación. Se hace notar que para la evaluación del test original se utilizó la información de 695 sustentantes que rindieron esta evaluación en los períodos 2020-I, 2020-II, 2021-I y 2021-II; mientras que para la evaluación del test corregido se utiliza la información de 100 sustentantes que rindieron en el período 2022-I. El instrumento de evaluación consta de 20 ítems con 4 opciones de respuesta cada uno, los cuales han sido codificados con 1 en caso de acertar a la respuesta y 0 en caso contrario; por lo que se considera, un test con resultados dicotómicos [8], [13]. Por lo tanto, los objetos de estudio son el instrumento de evaluación (test) y el grupo de sustentantes mencionados previamente.

El test de matemática se enfocó en las temáticas: aritmética, álgebra básica, sistemas de ecuaciones, funciones y trigonometría. La técnica utilizada para la recopilación de datos fueron estos tests que se compilaron en un formulario de Google forms, el cual fue distribuido a cada uno de los sustentantes por medios electrónicos. El tiempo destinado para esta evaluación fue de 2 horas para cada uno de los sustentantes, en un mismo horario por período.

Para el análisis cuantitativo del instrumento de evaluación se utiliza el enfoque de la TRI, en particular se aplica el modelo matemático de Rasch; se elige este modelo ya que cumple los principios de un modelo de medición como: proporcionar medidas lineales en intervalos iguales, emitir estimaciones más precisas, detectar la imprecisión del modelo y facilitar instrumentos de medición independientes de los parámetros estudiados [8], [9], [11], [14].

Por lo tanto, con la primera base de datos del test de diagnóstico, se procedió a la ejecución del modelo de Rasch, utilizando la librería *eRm* en el software R versión 4.1.1 y obteniendo de esta forma la dificultad de cada ítem y la habilidad de cada sustentante. Posteriormente, se procede a obtener los indicadores de validez y confiabilidad del instrumento de medición, como son el alfa de Cronbach y los parámetros de Pearson, para lo cual se utilizó las librerías *psych, psychometric* y *psycho.* Con base en los parámetros de Pearson se identifica los ítems que nos son descritos por el modelo, considerando el criterio del

Figura 1.
Curva característica del ítem en el modelo de Rasch



Nota. Se presenta la CCI, donde β es la dificultad del ítem y θ la habilidad del sustentante. Fuente: Adaptado [9].

p valor menor a 0,05 y los siguientes criterios adicionales [8], [11], [15]:

1. X^2 toma valores grandes
2. La discriminación debe ser menor a 0,19
3. El Infit MSQ y el Outfit MSQ deben corresponder a valores fuera del intervalo (0,8: 1,2)

Adicionalmente, se categoriza a los sustentantes en función de la escala de habilidad propuesta en [11], de donde se identifica aquellos con escasa habilidad en matemática.

El análisis del test corregido se realiza siguiendo la misma metodología descrita previamente para el test original. Luego, se procede a realizar las comparaciones sobre los resultados emitidos por ambos tests y se obtiene la simulación de las notas de cada sustentante, procediendo finalmente a hacer un ANOVA entre las notas reales y las simuladas.

3. Resultados y discusión

3.1. RESULTADOS DE LOS ÍTEMS DE AMBOS TESTS

En las tablas 1 y 2, se presenta la columna «Media dificultad» que corresponde al promedio de la dificultad de cada ítem en función de las respuestas de los sustentantes. Además, la columna «Dificultad SN» corresponde a los valores obtenidos por el modelo de Rasch de la dificultad sin normalizar, de cada ítem en función de las respuestas de los sustentantes; mientras que la columna «Dificultad N» representa los resultados normalizados ($N(0,1)$) de la dificultad de cada ítem, emitido por el modelo (ver Tabla 1).

Es necesario comentar que la tabla 1 corresponde a los resultados del test original, mientras que la tabla 2 recoge los resultados del test corregido (ver Tabla 2).

En la tabla 3, se presenta los parámetros de Pearson entre ellos la suma de los mínimos cuadrados residuales MSQ para la distribución x^2, además del índice de discriminación y el p-valor. Se observa que los ítems 2, 3, 8, 11, 15, 19, y 20 verifican los criterios establecidos en la metodología para su reajuste (ver Tabla 3).

Además, la confiabilidad del test original es de 0,65, puesto que 7 de los 20 ítems no son bien descritos por el modelo; mientras que, el valor del alfa de Cronbach es de 0,645117. Es de notar que el alfa de Cronbach se consideró debido a que se tiene ítems dicotómicos [16].

En función de los resultados de la tabla 3 se procede a corregir los ítems 2, 3, 8, 11, 15, 19, y 20. Por lo que en la tabla 4 se presenta los parámetros de Pearson para el test corregido (ver Tabla 4).

La confiabilidad del test corregido en base al modelo de Rasch es de 0,90 puesto que 2 de los 20 ítems no son bien descritos; mientras que el valor del alfa de Cronbach es de 0,7213858.

Por otro lado, en la figura 2 se puede visualizar las curvas características de los ítems del test original, que son el resultado de la interpolación de los puntos calculados por la probabilidad de la habilidad del sustentante, con cada valor fijo de la dificultad del ítem. Note que, si $P(\theta_k)=1/2$, entonces la habilidad (θ_k) coincide con el valor de la dificultad del ítem (β_i) (ver Figura 2).

Así también, en la figura 3 se observa las curvas características de los ítems del test corregido, note que la CCI del ítem 14 se encuentra a la derecha en la figura 3, con una dificultad de 1,778862 (ver Figura 3).

En la figura 4 se presenta un diagrama de caja que corresponde a las notas del test original y las notas simuladas respectivas; mientras que la figura 5 representa a las notas del test corregido y las simuladas del mismo (ver Figura 4). Para estas simulaciones se utiliza la ecuación:

$$\text{Notas}=\omega(\textit{base original parametrizada con 0 y 1}\times\beta^t) \quad (10)$$

Tabla 1.
Valores de las dificultades (β) del modelo Rasch, del test original

Ítems	Media dificultad	Dificultad SN	Dificultad N
Ítem 1	0,6576	0,121	0,5481
Ítem 2	0,7439	0,5772	0,7181
Ítem 3	0,918	2,0205	0,9783
Ítem 4	0,5165	-0,5291	0,2984
Ítem 5	0,4763	-0,7074	0,2397
Ítem 6	0,3439	-1,3105	0,095
Ítem 7	0,5842	-0,2256	0,4108
Ítem 8	0,8561	1,3535	0,9121
Ítem 9	0,5237	-0,4972	0,3095
Ítem 10	0,5799	-0,2453	0,4031
Ítem 11	0,6058	-0,1262	0,4498
Ítem 12	0,2662	-1,706	0,044
Ítem 13	0,6504	0,0857	0,5342
Ítem 14	0,8417	1,2333	0,8913
Ítem 15	0,5612	-0,33	0,3707
Ítem 16	0,4647	-0,7584	0,2241
Ítem 17	0,4489	-0,8287	0,2036
Ítem 18	0,5022	-0,5928	0,2766
Ítem 19	0,8245	1,0998	0,8643
Ítem 20	0,8576	1,3661	0,914

Tabla 2.
Valores de las dificultades (β) del modelo Rasch, del test corregido

Ítems	Media dificultad	Dificultad SN	Dificultad N
Ítem 1	0,6	0,098856	0,539374
Ítem 2	0,61	0,146676	0,558306
Ítem 3	0,72	0,712994	0,762075
Ítem 4	0,33	-1,16114	0,122793
Ítem 5	0,38	-0,91762	0,179408
Ítem 6	0,36	-1,01351	0,155407
Ítem 7	0,45	-0,5922	0,276858
Ítem 8	0,74	0,827894	0,796135
Ítem 9	0,48	-0,45531	0,324444
Ítem 10	0,56	-0,08899	0,464547
Ítem 11	0,73	0,769812	0,779294
Ítem 12	0,63	0,243688	0,596264
Ítem 13	0,63	0,243688	0,596264
Ítem 14	0,87	1,778862	0,962369
Ítem 15	0,5	-0,36422	0,357846
Ítem 16	0,45	-0,5922	0,276858
Ítem 17	0,52	-0,27298	0,392434
Ítem 18	0,49	-0,40977	0,340988
Ítem 19	0,69	0,549103	0,708533
Ítem 20	0,68	0,496367	0,690182

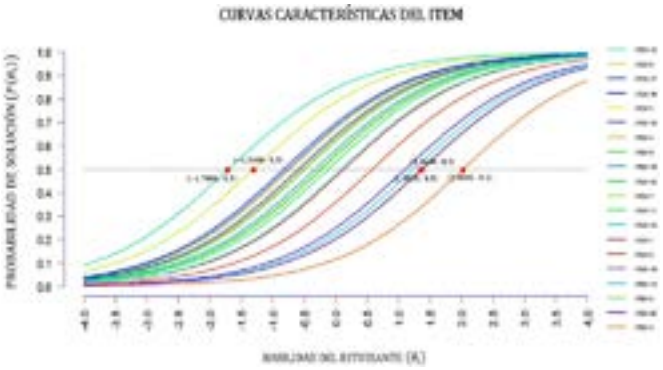
Tabla 3.
Valores de los parámetros de Pearson del modelo Rasch del test de diagnóstico

Ítems		Outfit MSQ	Infit MSQ	Outfit t	Infit t	Discriminación	p valor
Ítem 1	691,0291	0,9943	0,9794	-0,097	-0,5731	0,3237	0,525
Ítem 2	893,7067	1,2859	1,1132	3,8608	2,4349	0,0901	0
Ítem 3	812,284	1,1688	0,9623	1,0386	-0,3358	0,1251	0,001
Ítem 4	755,3017	1,0868	1,0538	2,2657	1,8743	0,2068	0,053
Ítem 5	560,5039	0,8065	0,8191	-5,4936	-6,878	0,5657	1
Ítem 6	543,7068	0,7823	0,8506	-4,6623	-4,7679	0,4746	1
Ítem 7	655,2027	0,9427	0,9558	-1,4085	-1,4597	0,3728	0,852
Ítem 8	899,1171	1,2937	1,0238	2,4522	0,3594	0,0811	0
Ítem 9	635,1644	0,9139	0,9136	-2,3469	-3,1192	0,4285	0,946
Ítem 10	692,5392	0,9965	1,0068	-0,0736	0,2331	0,2959	0,509
Ítem 11	814,8758	1,1725	1,145	3,795	4,3775	0,0948	0,001
Ítem 12	607,8396	0,8746	0,8931	-1,9722	-2,7057	0,3456	0,992
Ítem 13	559,5062	0,805	0,8091	-4,2385	-5,8615	0,5847	1
Ítem 14	524,5884	0,7548	0,8654	-2,5566	-2,1257	0,413	1
Ítem 15	778,1422	1,1196	1,0955	2,9451	3,1438	0,1685	0,014
Ítem 16	699,9882	1,0072	1,0122	0,2007	0,4404	0,2749	0,429
Ítem 17	644,932	0,928	0,955	-1,8863	-1,6053	0,3647	0,908
Ítem 18	711,6753	1,024	1,0242	0,6502	0,8623	0,2615	0,313
Ítem 19	896,2762	1,2896	1,0893	2,81	1,4465	0,0634	0
Ítem 20	775,6238	1,116	1,0104	1,0341	0,17	0,1557	0,017

Tabla 4.
Valores de los parámetros de Pearson del modelo Rasch del test corregido

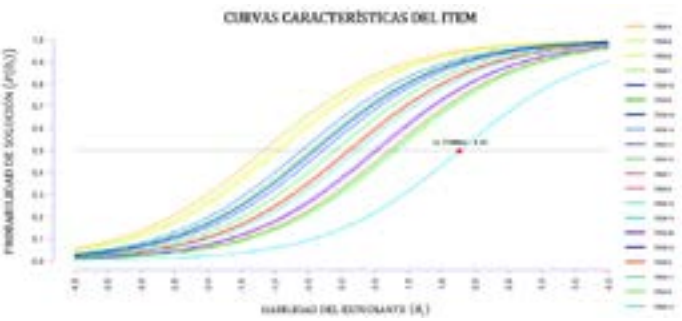
Ítems		Outfit MSQ	Infit MSQ	Outfit t	Infit t	Discriminación	p valor
Ítem 1	103,8782	1,0388	1,0708	0,3448	0,8255	0,2759	0,349
Ítem 2	146,6174	1,4662	1,2798	3,1976	2,9415	0,0335	0,001
Ítem 3	100,4332	1,0043	1,0563	0,0840	0,5241	0,2630	0,441
Ítem 4	63,6830	0,6368	0,7385	-2,5789	-3,0897	0,6335	0,998
Ítem 5	71,5454	0,7155	0,8034	-2,2494	-2,4736	0,5790	0,983
Ítem 6	71,6056	0,7161	0,7931	-2,1206	-2,5308	0,6039	0,983
Ítem 7	107,5877	1,0759	1,0988	0,6602	1,2361	0,2303	0,261
Ítem 8	122,7719	1,2277	1,0811	1,1621	0,6938	0,1828	0,053
Ítem 9	89,2422	0,8924	0,9454	-0,9291	-0,6810	0,4194	0,749
Ítem 10	104,6747	1,0467	1,0231	0,4293	0,3064	0,3221	0,329
Ítem 11	101,3890	1,0139	1,0332	0,1363	0,3191	0,2714	0,415
Ítem 12	107,3591	1,0736	1,0560	0,5740	0,6313	0,2925	0,266
Ítem 13	68,3148	0,6831	0,7435	-2,5874	-3,0527	0,6721	0,992
Ítem 14	119,9203	1,1992	1,0659	0,6613	0,3853	0,1471	0,075
Ítem 15	79,6080	0,7961	0,7988	-1,8835	-2,7143	0,6012	0,924
Ítem 16	95,6323	0,9563	0,9930	-0,3317	-0,0619	0,3628	0,577
Ítem 17	82,0540	0,8205	0,8546	-1,6360	-1,8950	0,5340	0,891
Ítem 18	113,6111	1,1361	1,1453	1,1747	1,7880	0,2104	0,15
Ítem 19	148,7692	1,4877	1,0868	2,6680	0,8410	0,2007	0,001
Ítem 20	113,1220	1,1312	1,1486	0,8525	1,4226	0,1525	0,157

Figura 2.
Curvas características de los ítems del test original, usando el modelo de Rasch.



Nota. Representa las curvas características de los 20 ítems del test original, el valor 0,5 corresponde a la dificultad media y 0,0 es la habilidad media. Fuente: Elaboración Propia

Figura 3.
Curvas características de los ítems del test corregido, en el modelo de Rasch



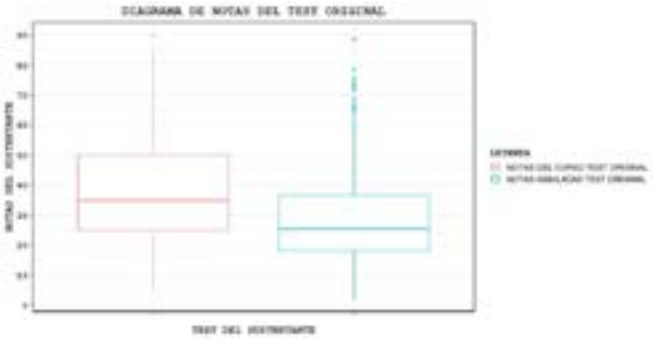
Nota. Representa las curvas características de los 20 ítems del test corregido, el ítem 13 presenta la mayor dificultad. Fuente: Elaboración propia

donde β es el vector de las dificultades de los ítems y ω se obtiene mediante la siguiente fórmula,

$$\omega = \frac{100}{\left(\sum_{i=1}^{n} \beta_i\right)} \qquad (11)$$

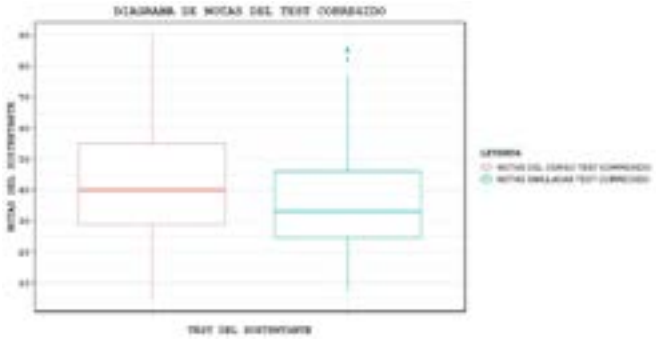
Se aclara que en la figura 4 se recoge los resultados de los 695 sustentantes de los períodos 2020-I hasta 2021-II, mientras que en la figura 5 se presenta los resultados de los 100 sustentantes que rindieron el test corregido en el período 2022-I (ver Figura 5).

Figura 4.
Diagrama de caja de las notas reales del test original y las notas simuladas



Nota. Corresponde a las notas reales (en rosa) y simuladas (en azul) del test original, las medias son diferentes. Fuente: Elaboración propia

Figura 5.
Diagrama de caja de las notas reales del test corregido y las notas simuladas



Nota. Corresponde a las notas reales (en rosa) y simuladas (en azul) del test corregido, las medias son diferentes. Fuente: Elaboración propia

3.2. RESULTADOS DE LOS SUSTENTANTES DE AMBOS TESTS

En efecto, como resultados adicionales del modelo de Rasch se obtiene que la confiabilidad de la habilidad de los sustentantes, que fueron evaluados con el test original, es de 0,80; debido a que 140 de los 695 no son descritos por el modelo, según el indicador de 0,19 de los parámetros de Pearson, mencionado en la metodología.

Asimismo, el modelo arroja una confiabilidad de 0,74 correspondiente a la habilidad de los sustentantes, que fueron evaluados con el test corregido, esto se debe a que 36 de los 100 no son descritos por el modelo de Rasch.

En [11] se propone el rango de habilidad siguiente: si la habilidad es menor a -1, entonces se considera que el sustentante tiene baja habilidad; si la habilidad se encuentra entre los valores de -1 a 1, entonces se considera una habilidad media o moderada; mientras que se considera una alta habilidad si es mayor a 1.

En consecuencia, se observa que la figura 6 recoge en rangos las habilidades de los 695 sustentantes presentados en la figura 4. Por lo que se nota que 76 de los 695 sustentantes caen en un rango de baja habilidad; 523 sustentantes se consideran con habilidad moderada, mientras que 96 sustentantes tienen una alta habilidad en matemática (ver Figura 6).

En la figura 7 se categoriza los resultados visualizados previamente en la figura 5, considerando las habilidades de los sustentantes en los rangos de habilidad baja, media y alta. Por lo tanto 14 de los 100 sustentantes se clasifican con baja habilidad; 68 se clasifican con habilidad media; mientras que 18 presentan una alta habilidad en matemática. Se ha considerado el mismo rango propuesto en [11] (ver Figura 7).

Adicionalmente, en la tabla 5 se presenta los resultados del ANOVA, entre las notas reales y simuladas tanto del test original como del test corregido (ver Tabla 5).

Finalmente, en la figura 8 se observa el diagrama de cajas de las diferencias entre las notas reales y simuladas, tanto del test original, como del test corregido; se procede a realizar una prueba de hipótesis de una cola, en la cual, la hipótesis nula sostiene que la media de la diferencia del test corregido es menor o igual a la media de la diferencial del test original, obteniéndose un p-valor de 3,07e-09 (ver Figura 8).

Discusión

En función de los resultados de la tabla 1 se sigue que los ítems 3, 8 y 20 corresponden a los de mayor dificultad del test original con índices de dificultad 0,9783; 0,9121 y 0,9140, respectivamente, que en comparación con los resultados del test corregido, este último presenta solamente un ítem de elevada dificultad (ítem 14), como se observa en la tabla 2. Asimismo, los ítems 6 y 12 de la tabla 1 poseen un sesgo mínimo por lo que corresponden a los ítems de menor dificultad en el test original, con índices de dificultad correspondientes 0,0950 y 0,0440; lo cual en comparación con la tabla 2, no existen ítems con dificultad menores a 0,1. Además, 0,9912618 es la correlación entre las variables *Media Dificultad* y *Dificultad N* del test original; por otro lado la correlación de las mismas variables en el test corregido es de 0,997475, de donde se observa un incremento en la correlación del test corregido respecto al test original con un nivel de confianza del 95%.

Por lo tanto, el modelo de Rasch a partir de los resultados presentes en la tabla 3, sugiere realizar un ajuste a los ítems 2, 3, 8, 11, 15, 19 y 20; esto puede deberse a factores como problemas de redacción, mala estructura de la pregunta del ítem, entre otras dificultades en la elaboración de los mismos [17]-[19]. Por lo tanto, luego de

Tabla 5.
ANOVA de las notas reales y simuladas de ambos tests

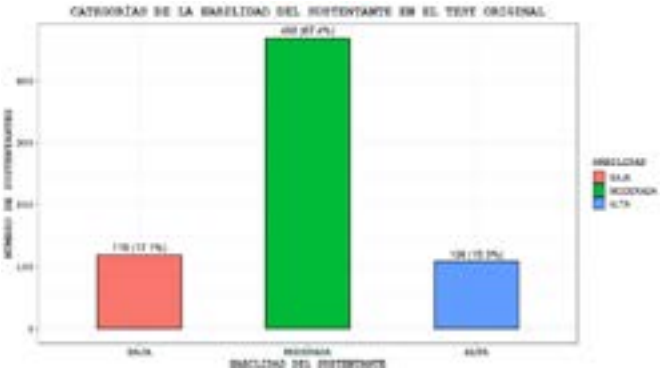
	Df	Sum Sq	Mean Sq	F value	Pr(>F)
Notas del test original	1	55814	55814	218	<2e-16
Notas del test corregido	1	4632	4632	12,83	0,00043

Tabla 6.
Rango de habilidad para el alfa de Cronbach

Rango	Interpretación
	Excelente
	Bueno
	Aceptable
	Pobre
	Inaceptable

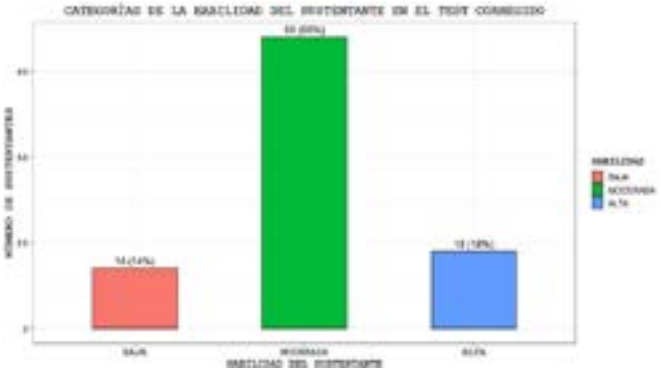
Nota. Adaptado [20].

Figura 6.
Rangos de habilidad de los sustentantes que rindieron el test original



Nota. Se observa las frecuencias de la habilidad en matemática de los sustentantes, categorizadas en baja, moderada y alta del test original. Fuente: Elaboración propia

Figura 7.
Rangos de habilidad de los sustentantes que rindieron el test corregido



Nota. Se observa las frecuencias de la habilidad en matemática de los sustentantes, categorizadas en baja, moderada y alta del test corregido. Fuente: Elaboración propia

Figura 8.
Diagrama de caja de la diferencia de notas de ambos tests

Nota. Se observa que la media de la diferencia de notas del test original es ligeramente superior a la media de la diferencia de las notas del test corregido. Fuente: Elaboración propia

realizadas las correcciones de los ítems indicados previamente, se visualiza en la tabla 4, un notable ajuste de los ítems al modelo de Rasch; esto corroborado por un incremento de la confiabilidad del instrumento de evaluación de 0,65 presente en el test inicial al 0,9 presente en el test corregido; además el alfa de Cronbach con un valor de 0,645117 que representa un indicador aceptable pasa a un valor 0,7213858, que entra en un rango bueno, tal como se muestra en la tabla 6 obtenida de [20] (ver tabla 6).

Por otro lado, del ANOVA resumido en la tabla 6 se observa un primer p-valor correspondiente a 2e-16, lo cual implica el rechazo de la hipótesis nula correspondiente a la igualdad de medias, es decir, hay evidencia estadística que confirma que las medias de las notas reales del test

original y la media de las notas simuladas son diferentes. Así mismo, para el test corregido se observa un p-valor de 0,00043 por lo que existe suficiente evidencia para asegurar que las medias de las notas reales y simuladas son diferentes.

Sin embargo, a pesar de que no se verifica las igualdades de las medias, las diferencias de las notas reales con respecto a las notas simuladas de ambos test, disminuye como se muestra en la figura 8, donde al realizar una prueba de hipótesis de una cola, se obtiene el p-valor de 3,074e-09, por lo que se rechaza la hipótesis nula, es decir que existe evidencia estadística suficiente para afirmar que la media de la diferencia entre las notas verdaderas y simuladas del test original es mayor que la media de la diferencia entre las notas reales y simuladas del test corregido. Se observa además en la figuras 8, que la media de la diferencia de las notas del test original resulta 10,44355; mientras que la media de la diferencia de las notas del test corregido corresponde a 6,63685, lo cual representa una mejora del 36,45% en el ajuste de la simulación, posterior a la corrección del test original.

Asimismo, la correlación entre las notas simuladas y las reales del test original es de 0,9156131; mientras que la correlación en el test corregido corresponde a 0,9567291; evidenciando también un incremento. En consecuencia, hay evidencia para afirmar que las notas simuladas del test corregido, se ajusta de mejor forma a las notas reales de los sustentantes, en comparación con las notas simuladas con el test original.

En las figuras 2 y 3 se presenta las curvas características de los ítems, tanto del test original como del corregido, respectivamente, donde se observa que las CCI del test corregido se agrupan más uniformemente que las del test original, obteniendo de esta forma un instrumento de evaluación que describe con mayor precisión el modelo. Así, entonces, el test corregido representa el instrumento de evaluación válido y confiable buscado.

Como resultados de las figuras 4 y 6 se observa que el 17,1% de los sustentantes que rindieron el test original tienen una habilidad baja, este porcentaje corresponde a 119 sustentantes. Así también, en las figuras 5 y 7 se obtiene el 14% de sustentantes con un bajo nivel de habilidad en matemática que rindieron el test corregido, este valor corresponde a 14 sustentantes. En consecuencia, se ha encontrado en ambos test, en total 133 sustentantes con habilidad baja. Adicionalmente, se observa que la confiabilidad de la habilidad de los sustentantes que rindieron el test original es de 80%, mientras que la confiabilidad de los que rindieron el test corregido es de 74%, esta disminución en la confiabilidad se debe a la gran disminución de los sustentantes que rindieron este último test respecto al original.

En este trabajo se ha encontrado evidencia suficiente de la fiabilidad en la aplicación del modelo de Rasch, en comparación con las investigaciones de [8], [11].

4. CONCLUSIÓN

Este estudio ha permitido avanzar en el desarrollo de la compleja relación entre la creación de un instrumento de evaluación, independiente de la habilidad de los individuos que son evaluados con dicho instrumento. [21], [22]

La utilización del modelo de Rasch ha permitido determinar a los individuos con deficientes habilidades en matemática, cuya información se ha facilitado a las coordinaciones pertinentes, para que a su vez realicen el seguimiento necesario y se logre disminuir en lo posible el nivel de deserción estudiantil, lo cual consiste en un problema muy recurrente en las instituciones de educación superior.

Por otro lado, en este trabajo se ha codificado los resultados de los test de forma dicotómica, como 1 en caso de acertar y 0 en caso de no acertar al ítem; sin embargo, cada ítem tiene cuatro opciones de respuesta, por lo que en otro estudio se podría considerar un caso no bidimensional y determinar si estas opciones de solución afectan significativamente a la habilidad del sustentante o a la dificultad del test. Asimismo, se puede realizar un análisis utilizando un modelo de Rasch más preciso, es decir, de dos o tres parámetros; sin embargo, para estos modelos se requiere considerar una mayor cantidad de datos, lo cual beneficiaría en un mejor ajuste de la simulación de las notas.

REFERENCIAS

[1] ISUCT, *Informe de gestión administrativa y académica-Rendición de cuentas 2020*, 2020, p. 105. [Online]. Available: https://istct.edu.ec/portal/nuevo/wp-content/uploads/sites/2/2021/06/Rendición-de-cuentas-2020-V002.pdf

[2] T. M. Bechger, G. Maris, H. H. F. M. Verstralen, and A. A. Béguin, «Using classical test theory in combination with item response theory», *Appl Psychol Meas*, vol. 27, N.º 5, pp. 319-334, sep. 2003, DOI: 10.1177/0146621603257518.

[3] J. Muñiz, «Las teoría de los tests: TCT y TRI», *Papeles del Psicólogo*, vol. 31, N.º 1, pp. 57-66, 2010, [Online]. Available: http://www.papelesdelpsicologo.es/pdf/1796.pdf

[4] R. L. Brennan, «Generalizability theory and classical test theory», *Applied Measurement in Education*, vol. 24, N.º 1, pp. 1-21, jan. 2011, DOI: 10.1080/08957347.2011.532417.

[5] J. M. Francese Martínez, *Medición y evaluación educativa*, 1st ed., Arco Libros - La Muralla, S. L., 2008.

[6] G. H. Fischer and I. W. Molenaar, *Rasch Models: Foundations, Recent Developments, and Applications.* New York: Springer-Verlag, 1995. DOI: 10.1007/978-1-4612-4230-7.

[7] N. Cortada de Kohan, «Teoría de respuesta al ítem: supuestos básicos», *Revista Evaluar*, vol. 4, N.º 1, pp. 95-110, 2004, DOI: 10.35670/1667-4545.v4.n1.600.

[8] K. Jiménez Alfaro y E. Montero Rojas, «Aplicación del

modelo de Rasch, en el análisis psicométrico de una prueba de diagnóstico en matemática», *Revista Digital: Matemática, Educación e Internet*, vol. 13, n.º 1, pp. 1-24, 2013, DOI: 10.18845/rdmei.v13i1.1628.

[9] I. Leenen, «Virtudes y limitaciones de la teoría de respuesta al ítem para la evaluación educativa en las ciencias médicas», *Investigación en Educación Médica*, vol. 3, n.º 9, pp. 40-55, 2014, DOI: 10.1016/s2007-5057(14)72724-3.

[10] E. San Martín, «Modelos Rasch: ¿cuán (in-)coherentemente son presentados y utilizados?», *Actualidades en Psicología*, vol. 29, n.º 119, pp. 91-102, 2015, DOI: http://dx.doi.org/10.15517ap.v29i119.18911.

[11] A. Atikah, S. Sudiyatno, A. Rahim, and M. Marlina, «Assessing the item of final assessment mathematics test of junior high school using Rasch model», *Jurnal Elemen*, vol. 8, n.º 1, pp. 117-130, 2022, DOI: 10.29408/jel.v8i1.4482

[12] A. Rahim and H. Haryanto, «Implementation of item response theory (IRT) Rasch model in quality analysis of final exam tests in mathematics», *Journal of Educational Research and Evaluation*, vol. 10, n.º 2, pp. 57-65, 2021, DOI: 10.15294/jere.v10i2.51802.

[13] S. Rakkapao, S. Prasitpong, and K. Arayathanitkul, «Analysis test of understanding of vectors with the three-parameter logistic model of item response theory and item response curves technique», *Phys Rev Phys Educ Res*, vol. 12, n.º 2, pp. 1-10, 2016, DOI: 10.1103/PhysRevPhysEducRes.12.020135.

[14] T. Verguts, P. De Boeck, and G. Storms, «Analyzing experimental data using the Rasch model», *Behavior Research Methods, Instruments, and Computers*, vol. 30, n.º 3, pp. 501-505, 1998, DOI: 10.3758/BF03200683.

[15] K. S. Sidhu, *New approaches to measurement and evaluation*. Sterling Publishers Pvt. Ltd., 2005.

[16] A. Freiberg Hoffmann, J. B. Stover, G. De la Iglesia and M. Fernández Liporace, «Correlaciones policóricas y tetracóricas en estudios factoriales exploratorios y confirmatorios», *Ciencias Psicológicas*, vol. 7, n.º 2, pp. 151-164, 2013, DOI: 10.22235/cp.v7i1.1057.

[17] T. Nielsen, «The specific academic learning self-efficacy and the specific academic exam self-efficacy scales: construct and criterion validity revisited using Rasch models», *Cogent Education*, vol. 7, n.º 1, 2020, DOI: 10.1080/2331186X.2020.1840009.

[18] F. Flores, M. Sánchez y A. Martínez, «Modelo de predicción del rendimiento académico de los estudiantes del ciclo básico de la carrera de Medicina a partir de la evaluación del desempeño docente», *Revista Mexicana de Investigación Educativa*, vol. 21, n.º 70, pp. 975-991, 2016, [Online]. Available: https://www.redalyc.org/pdf/140/14046162015.pdf

[19] S. Celis, L. Moreno, P. Poblete, J. Villanueva y R. Weber, «Un modelo analítico para la predicción del rendimiento académico de estudiantes de ingeniería», *Revista Ingeniería de Sistemas*, n.º septiembre 2015, pp. 5-24, 2015, [Online]. Available: https://www.researchgate.net/publication/292982515_Un_modelo_analitico_para_la_prediccion_del_rendimiento_academico_de_estudiantes_de_ingenieria

[20] M. Muntazhimah and R. Wahyuni, «The development and validation of mathematical reflective thinking test for prospective mathematics teachers using the Rasch model», *Jurnal Elemen*, vol. 8, n.º 1, pp. 175-186, 2022, DOI: 10.29408/jel.v8i1.3981.

[21] T. G. Bond, Z. Yan and M. Heene, *Applying the rasch model-fundamental measurement in the human sciences*, 4th ed., New York and London: Routledge, 2013. DOI: 10.4324/9781410614575.

[22] E. Backhoff, M. J. González Montesinos, Y. Pérez Garibay, and M. F. Ferreyra, «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*, vol. 20, n.º 1, pp. 41-55, 2022, DOI: 10.15366/reice2022.20.1.003.