

**LATAM Revista Latinoamericana de Ciencias Sociales y
Humanidades, Asunción, Paraguay.**

ISSN en línea: 2789-3855, 2025, Volumen VI

Clasificación de ítems de pruebas estandarizadas mediante conjuntos independientes maximales: Un enfoque basado en grafos, Teoría de Respuesta al Ítem y Teoría Clásica de Test

Classification of standardized test items by maximal
independent sets: A graph-based approach, Item Response
Theory and Classical Test Theory

Roberto Contreras Juárez

roberto.contreras@correo.buap.mx
<https://orcid.org/0000-0002-3271-6754>
Benemérita Universidad Autónoma de
Puebla
Puebla – México

Etelvina Archundia Sierra

etelvina.archundia@correo.buap.mx
<https://orcid.org/0000-0001-9686-5305>
Benemérita Universidad Autónoma de
Puebla
Puebla – México

Héctor David Ramírez Hernández

hector.ramirezhe@correo.buap.mx
<https://orcid.org/0000-0003-3741-4285>
Benemérita Universidad Autónoma de
Puebla
Puebla – México

Nelva Betzabel Hernández Espinosa

nelva.espinosa@correo.buap.mx
<https://orcid.org/0000-0002-5620-2336>
Benemérita Universidad Autónoma de
Puebla
Puebla – México

Mariana Ayala Munguía

mariana.ayalam@alumno.buap.mx
<https://orcid.org/0009-0003-0176-2496>
Benemérita Universidad Autónoma de
Puebla
Puebla – México

DOI: <https://doi.org/10.56712/latam.v6i4.4338>

Artículo recibido: 18 de junio de 2025

Aceptado para publicación: 15 de agosto de
2025.

Conflictos de Interés: Ninguno que declarar.


Redilat
Red de Investigadores
Latinoamericanos

NÚMERO

DOI: <https://doi.org/10.56712/latam.v6i4.4338>

Clasificación de ítems de pruebas estandarizadas mediante conjuntos independientes maximales: Un enfoque basado en grafos, Teoría de Respuesta al Ítem y Teoría Clásica de Test

Classification of standardized test items by maximal independent sets: A graph-based approach, Item Response Theory and Classical Test Theory

Roberto Contreras Juárez¹

roberto.contreras@correo.buap.mx

<https://orcid.org/0000-0002-3271-6754>

Benemérita Universidad Autónoma de Puebla
Puebla – México

Etelvina Archundia Sierra

etelvina.archundia@correo.buap.mx

<https://orcid.org/0000-0001-9686-5305>

Benemérita Universidad Autónoma de Puebla
Puebla – México

Héctor David Ramírez Hernández

hector.ramirezhe@correo.buap.mx

<https://orcid.org/0000-0003-3741-4285>

Benemérita Universidad Autónoma de Puebla
Puebla – México

Nelva Betzabel Hernández Espinosa

nelva.espinosa@correo.buap.mx

<https://orcid.org/0000-0002-5620-2336>

Benemérita Universidad Autónoma de Puebla
Puebla – México

Mariana Ayala Munguía

mariana.ayalam@alumno.buap.mx

<https://orcid.org/0009-0003-0176-2496>

Benemérita Universidad Autónoma de Puebla
Puebla – México

Artículo recibido: 18 de junio de 2025. Aceptado para publicación: 15 de agosto de 2025.
Conflictos de Interés: Ninguno que declarar.

Resumen

Este trabajo propone una metodología innovadora para la clasificación de ítems en pruebas estandarizadas, integrando la Teoría de Respuesta al Ítem (TRI) y la Teoría Clásica de los Test (TCT) con herramientas de la teoría de grafos. A partir de parámetros psicométricos como la dificultad (TRI), la discriminación (TCT) y la proporción de aciertos por ítem (p -value), se construye un grafo donde cada nodo representa un ítem y las aristas se trazan según umbrales de similitud establecidos entre dichos parámetros. Sobre esta estructura, se aplican algoritmos para extraer conjuntos independientes maximales (MIS), los cuales agrupan ítems suficientemente distintos entre sí en términos psicométricos. Estos MIS permiten una clasificación representativa, evitando redundancia y favoreciendo la cobertura de la prueba en distintos niveles de dificultad y discriminación. El enfoque facilita la identificación de tipologías de ítems y puede ser extendido para validar la estructura de la prueba mediante técnicas de agrupamiento y visualización. Los resultados preliminares con datos tomados de una prueba estandarizada muestran que esta estrategia permite un análisis más

estructurado y diverso de los ítems, promoviendo una mejora en el diseño y evaluación de instrumentos educativos. Se sugiere como una herramienta complementaria para investigadores y diseñadores de pruebas interesados en enfoques interdisciplinarios que integren estadística, psicometría y modelado computacional.

Palabras clave: conjunto independiente maximal, psicometría de ítems, clasificación, agrupamiento, pruebas estandarizadas, teoría de grafos

Abstract

This study proposes an innovative methodology for classifying items in standardized tests by integrating Item Response Theory (IRT) and Classical Test Theory (CTT) with graph theory tools. Using psychometric parameters such as difficulty (IRT), discrimination (CTT), and the proportion of correct responses per item (p -value), a graph is constructed where each node represents an item, and edges are drawn based on predefined thresholds of similarity among these parameters. Algorithms are then applied to extract maximal independent sets (MIS), which group items that are sufficiently different from each other in psychometric terms. These MIS enable a representative classification that avoids redundancy and promotes coverage across varying levels of difficulty and discrimination. This approach facilitates the identification of item typologies and can be extended to validate test structure through clustering and visualization techniques. Preliminary results with data taken from a standardized test show that the strategy supports a more structured and diversified item analysis, contributing to improved design and evaluation of educational assessments. The methodology is presented as a complementary tool for researchers and test designers interested in interdisciplinary frameworks that merge statistics, psychometrics, and computational modeling.

Keywords: maximal independent set, item psychometrics, classification, clustering, standardised tests, graph theory

Todo el contenido de LATAM Revista Latinoamericana de Ciencias Sociales y Humanidades, publicado en este sitio está disponibles bajo Licencia Creative Commons.



Cómo citar: Contreras Juárez, R., Archundia Sierra, E., Ramírez Hernández, H. D., Hernández Espinosa, N. B., & Ayala Munguía, M. (2025). Clasificación de ítems de pruebas estandarizadas mediante conjuntos independientes maximales: Un enfoque basado en grafos, Teoría de Respuesta al Ítem y Teoría Clásica de Test. *LATAM Revista Latinoamericana de Ciencias Sociales y Humanidades* 6 (4), 885 – 899. <https://doi.org/10.56712/latam.v6i4.4338>

INTRODUCCIÓN

En las pruebas estandarizadas es fundamental determinar la calidad del instrumento de medición (constructo). Esto conlleva a una vigilancia estricta del comportamiento de los ítems mediante el uso de distintos parámetros estadísticos, sin dejar de lado el análisis cualitativo que la propia medición del constructo indica, (Haladyna y Rodríguez, 2013).

Existen tres razones fundamentales por las que el análisis de ítems es de suma importancia, la primera es que permite identificar aquellos ítems que tienen un excelente desempeño, la segunda es que ayuda a mejorar los ítems con la finalidad de poder ser utilizados en pruebas nuevas y la tercera razón es que permite eliminar aquellos ítems que presentan problemas de construcción o que resultan ser ambiguos para ser eliminados.

En general, se pueden considerar tres etapas por las que un ítem nuevo debe pasar, a saber,

construcción del ítem; se lleva a cabo de acuerdo con las especificaciones previamente establecidas para vigilar la pertinencia del ítem en relación con el constructo intencionado, como los Standards for educational and psychological test, (AERA/APA/NCME, 2014).

revisión del ítem; este proceso incluye la participación de especialistas que verifican la pertinencia del ítem respecto del constructo, aquí se verifica el formato y se realizan las adecuaciones necesarias para garantizar el acceso a diferentes poblaciones que presentan la prueba, (Haladyna y Rodríguez, 2013; Moreno, et al, 2004).

evaluación del ítem; corresponde a un proceso en el cual se examinan las respuestas que los estudiantes ofrecen a cada uno de los ítems, permitiendo así, la evaluación de la calidad de dichos ítems y de la propia prueba en su conjunto, (Abad, et al, 2006; Ortiz, et al, 2015).

Como se puede apreciar, la evaluación de los ítems constituye una de las tareas más minuciosas y sensibles que se deben hacer antes de poder colocarlos en una prueba estandarizada, la cual es creada como medida para inferir las capacidades de las personas, ya que la función principal de un constructo, en el ámbito educativo, es ofrecer información de los examinados para la correcta toma de decisiones, (Backhoff, et al, 2000).

METODOLOGÍA

Para clasificar ítems usando conjuntos independientes máximos se tomó una muestra de 66 ítems de matemáticas de una prueba estandarizada que fue aplicada a una población de 16,256 estudiantes. Los ítems fueron evaluados a través de la TCT y de la TRI para obtener los parámetros p -value, r_{pbis} y b descritos con anterioridad y que son considerados en este trabajo.

Los parámetros psicométricos fueron obtenidos utilizando Xcalibre 4™, un software especializado en análisis psicométrico, diseñado para implementar la Teoría de Respuesta al Ítem (TRI) de forma automatizada y accesible. Su interfaz permite realizar calibraciones de ítems, análisis de distractores, detección de funcionamiento diferencial de ítems (DIF) y generación de informes narrativos con gráficos y tablas incrustadas, (van der Linden, 2016).

Los datos fueron recolectados en un archivo .csv donde se identificó a cada ítem mediante un número consecutivo. A cada ítem se le asignaron cuatro valores numéricos y un valor categórico, a saber,

ítem: es un número natural que sirve como identificador del ítem y corresponde al nodo dentro del grafo.

p_value: valor numérico correspondiente a la proporción de estudiantes que respondieron correctamente el ítem (p -value).

discriminacion: valor numérico correspondiente al índice de discriminación dado por el coeficiente de punto biserial r_{pbis} .

dificultad: valor numérico correspondiente al índice de dificultad obtenido a partir del parámetro b de la TRI.

tipo_item: clasificación del ítem de acuerdo con su índice de dificultad bajo las siguientes categorías:

Baja_Dific: Ítems de dificultad baja, aquellos que satisfacen $b < -1$.

Media_Dific: Ítems de dificultad media, aquellos que satisfacen $-1 \leq b \leq 1$.

Alta_Dific: Ítems de dificultad alta, aquellos que satisfacen $b > 1$.

La Tabla 2 muestra la estructura del archivo .csv y algunos ejemplos de los datos recolectados.

Tabla 2

Ejemplo de los datos contenidos en el archivo .csv.

item	dificultad	discriminacion	p_value	tipo_item
1	-2.1861	0.2609	0.7909	Baja_Dific
2	0.3220	0.3785	0.5926	Media_Dific
3	2.1151	0.3043	0.3211	Alta_Dific
4	2.108	0.2825	0.3853	Alta_Dific

En la modelación de los ítems como un grafo, cada nodo representa un ítem de la prueba. Para establecer la existencia de una arista entre dos ítems se definieron umbrales de similitud para el índice de discriminación y para la dificultad, definidos por la TCT y la TRI.

Umbral de similitud en discriminación (TCT): diferencia ≤ 0.10 entre los coeficientes de discriminación

Umbral de similitud en dificultad (TRI): diferencia ≤ 0.10 entre sus parámetros de dificultad .

Definidos los umbrales de similitud se establece la relación que determina la existencia de una arista entre dos nodos i y j , definida como

$$(i, j) \Leftrightarrow |b_i - b_j| \leq 0.10 \wedge |r_{pbis_i} - r_{pbis_j}| \leq 0.10$$

A partir de esta relación, se generó el código que se implementado en Python para crear el grafo bajo los umbrales de similitud requeridos.

Figura 1

Algoritmo en Python para crear el grafo

```

1  import pandas as pd
2  import networkx as nx
3  import matplotlib.pyplot as plt
4  import matplotlib.cm as cm
5  import matplotlib.colors as colors
6  from sklearn.cluster import KMeans
7  from sklearn.decomposition import PCA
8  from networkx.algorithms import approximation
9
10
11 # Reemplaza con la ruta a tu archivo
12 df = pd.read_csv("MIS_items.csv") |
13
14 # Convertir a diccionario para NetworkX
15 items = df.set_index('item').T.to_dict()
16
17 # Umbrales ajustables
18 umbral_dificultad = 0.1
19 umbral_discriminacion = 0.1
20
21 G = nx.Graph()
22 G.add_nodes_from(items.keys())
23
24 for i in items:
25     for j in items:
26         if i != j:
27             dif_b = abs(items[i]['dificultad'] - items[j]['dificultad'])
28             dif_rpbis = abs(items[i]['discriminacion'] - items[j]['discriminacion'])
29             if dif_b <= umbral_dificultad and dif_rpbis <= umbral_discriminacion:
30                 G.add_edge(i, j)

```

Esto ayuda a que un conjunto independiente maximal contenga ítems no conectados entre sí, es decir, suficientemente distintos en dificultad/discriminación y al ser maximal, no se puede agregar ningún otro ítem sin violar la independencia.

Por lo tanto, un MIS representará un conjunto diverso de ítems con diferencias suficientes para ser considerados distintos tipos de dificultad/discriminación, por lo que estos conjuntos pueden usarse como base representativa para diseñar o validar la estructura de una prueba.

DESARROLLO

La TCT y la TRI

Existen dos paradigmas psicométricos comúnmente utilizados en el análisis de ítems de pruebas estandarizadas: la Teoría Clásica de los Test (TCT) y la Teoría de la Respuesta al Ítem (TRI), estos dos paradigmas se utilizan para analizar el banco de reactivos que conforman el pretest.

La TCT representa una aproximación simple e intuitiva para el análisis de ítems y pruebas. Los ítems son descritos por proporciones y por correlaciones, mientras que los puntajes de los examinados corresponden a la suma total de los ítems que fueron respondidos correctamente. La TCT es muy útil para evaluar el funcionamiento de los ítems y sus respuestas individuales para trabajar en conjunto con los expertos en el contenido de la materia para mejorar sus propiedades, (Abad, et al, 2006).

La evaluación de ítems con TCT se enfoca en dos características de un ítem: dificultad y discriminación. El índice de dificultad en la TCT se denota como p y se calcula como la proporción de examinados que respondieron correctamente el ítem, es decir,

$$p = \frac{\text{número de aciertos}}{\text{Total de examinados}}$$

y sus valores oscilan entre 0.00 a 1.00.

A partir de los valores de p , es posible definir la escala de clasificación del nivel de dificultad de los ítems en un constructo. La escala de clasificación de la dificultad más frecuente de los ítems considera cinco intervalos, los cuales se describen en la Tabla 1, (Backhoff, et al, 2000; Ortiz, et al, 2015).

Tabla 1

Escala de clasificación de la dificultad usando p –value

Clasificación	Intervalo de p
Fácil	$0.90 < p \leq 1.00$
Medianamente fácil	$0.80 < p \leq 0.90$
Mediano	$0.50 < p \leq 0.80$
Medianamente difícil	$0.40 \leq p \leq 0.50$
Difícil	$0.00 \leq p < 0.40$

Para obtener el índice de discriminación de cada ítem, D_i , se calcula la diferencia entre la proporción de aciertos en el grupo superior y la proporción de aciertos en el grupo inferior, es decir,

$$D_i = \frac{GS_{aciertos} - GI_{aciertos}}{N_{max}}$$

donde:

D_i es el índice de discriminación del reactivo i .

$GS_{aciertos}$ es el número de aciertos del ítem i que el 27% de los examinados con las puntuaciones más altas obtuvieron en la prueba.

$GI_{aciertos}$ es el número de aciertos del ítem i que el 27% de los examinados con las puntuaciones más bajas obtuvieron en la prueba.

N_{max} es el número máximo de examinados pertenecientes al grupo superior o al grupo inferior, es decir, $N_{max} = \max \{GS_{aciertos}, GI_{aciertos}\}$.

Entre mayor sea D_i , el ítem es más discriminante, y mejor sitúa a un sujeto en el grupo superior o el grupo inferior, (Backhoff, et al, 2000; Morales 2012).

También es frecuente utilizar el coeficiente biserial (r_{bis}) y el punto biserial (r_{pbis}) para estimar el índice de discriminación. Ambas son aproximadamente proporcionales, sin embargo, el coeficiente biserial suele tener valores absolutos mayores, esto hace que el punto biserial sea mayormente utilizado, (Backhoff, et al, 2000).

$$r_{pbis} = \frac{X_{sup} - X_{inf}}{S_x} \sqrt{\frac{n_{sup}n_{inf}}{n(n-1)}}$$

donde

$\bar{X}_{sup}$ es la media de las puntuaciones de los examinados que respondieron correctamente el ítem.

$\bar{X}_{inf}$ es la media de las puntuaciones de los examinados que respondieron incorrectamente el ítem.

S_x es la desviación estándar de las puntuaciones totales obtenidas por los examinados.

n_{sup} es el número de respuestas correctas del ítem.

n_{inf} es el número de respuestas incorrectas del ítem.

El coeficiente punto biserial, (r_{pbis}), se utiliza para saber si los examinados “mejores” son los que obtienen las respuestas correctas, qué tanto poder predictivo tiene el reactivo y cómo éste puede contribuir a las predicciones, (Backhoff, et al, 2000). Esto lo convierte en una herramienta valiosa para evaluar la calidad de los ítems, ayudando a identificar aquellos que contribuyen eficazmente a la capacidad de la prueba para diferenciar entre los examinados.

Los valores de r_{pbis} varían entre -1.00 y $+1.00$. La interpretación de estos valores para la discriminación del ítem se resume generalmente en las siguientes categorías, (Muñiz, 2010):

$r_{pbis} > 0.35$ excelente poder de discriminación. El ítem discrimina muy bien entre los examinados con altas y bajas puntuaciones en la prueba total. Aquellos que rinden bien en la prueba general tienden a responder este ítem correctamente, y viceversa.

$0.25 < r_{pbis} \leq 0.35$ buen poder discriminativo. El ítem tiene una buena capacidad para diferenciar entre los niveles de habilidad. Son ítems que generalmente se consideran aceptables.

$0.15 < r_{pbis} \leq 0.25$ discriminación regular/aceptable. El ítem discrimina, pero su capacidad para hacerlo no es óptima. Podrían requerir revisión o mejora.

$0.0 \leq r_{pbis} \leq 0.15$ pobre discriminación/calidad marginal. El ítem tiene poca o ninguna capacidad para diferenciar entre los examinados. Sugieren que el ítem no está funcionando como se espera y probablemente deba ser revisado o reemplazado.

$r_{pbis} < 0.0$ discriminación negativa/Ítem malo. Un valor negativo indica que los examinados con puntuaciones bajas en la prueba total tienden a responder el ítem correctamente, y los examinados con puntuaciones altas en la prueba total tienden a responder incorrectamente.

Una discriminación negativa es un indicio de un problema grave con el ítem, por ejemplo, la clave de respuesta es incorrecta, el ítem está formulado de manera confusa o mide algo diferente a lo que mide el resto de la prueba. Estos ítems deben ser descartados o revisados a fondo.

Por su parte, la TRI es una aproximación más compleja y no provee información de las respuestas individuales, sin embargo, tiene una serie de ventajas en el diseño de formas paralelas, equiparación y ensamblaje de nuevas pruebas. Una ventaja adicional es que se pueden diseñar test adaptativos computarizados, (Weist & Kingsbury, 1984).

La estructura principal de la TRI son los parámetros de los ítems, cada ítem tiene tres parámetros asociados que se deben estimar, a saber,

Índice de discriminación α , este valor indica que tan bien el ítem puede diferenciar entre examinados con bajo y alto puntaje. Típicamente, el rango varía entre 0.00 a 2.00 , donde los valores más altos sugieren que el ítem discrimina mejor.

Índice de dificultad b , este valor indica para que tipo de examinado es mejor el ítem. Sus valores varían entre -3.00 a 3.00 , siendo 0.00 el nivel promedio de dificultad.

Índice de adivinación c , corresponde a la probabilidad de que un estudiante pueda acertar el ítem al azar y su valor es $c = \frac{1}{k}$ donde k es el número de opciones de respuesta del ítem.

Los valores de los parámetros se calculan a partir de los datos obtenidos al aplicar los ítems a una muestra amplia y representativa de personas. Para estos cálculos son necesarios sofisticados programas, razón por la cual los modelos de TRI se extendieron hasta que se dispuso de ordenadores potentes, (Muñiz, 2010).

La TRI es una familia de modelos psicométricos, en algunos casos se utiliza únicamente el parámetro b , mejor conocido como modelo de Rasch; en otros casos se utilizan los parámetros a y b . Sin embargo, para ítems de opción múltiple lo más recomendable es utilizar un modelo que estime los tres parámetros, (Embretson & Reise, 2000).

La fórmula del modelo logístico de tres parámetros es dada por

$$P(\theta) = c + (1 - c) \left[\frac{e^{1.7a(\theta-b)}}{1 + e^{1.7a(\theta-b)}} \right]$$

donde

$P(\theta)$ es la probabilidad de responder correctamente el ítem.

θ es la puntuación en la variable medida y toma valores entre -3.00 y $+3.00$.

a, b, c son los tres parámetros descritos anteriormente.

El valor θ (Theta) representa el nivel del rasgo latente de una persona, por ejemplo, habilidad, conocimiento o cualquier otro rasgo, pero también tiene implicaciones importantes para la interpretación de los ítems. Cada ítem en la TRI tiene una Curva Característica del Ítem (CCI) que muestra la probabilidad de que una persona con cierto nivel de θ responda correctamente. El eje horizontal representa los valores de θ y el eje vertical representa la probabilidad de responder correctamente al ítem. En términos prácticos, si un ítem tiene dificultad b , la probabilidad de que $\theta = b$ es 0.5 , (Abad, et al, 2006).

Conjuntos Independientes de un Grafo

La teoría de grafos ofrece un marco potente e intuitivo para modelar sistemas complejos, representando entidades como nodos y sus relaciones como aristas. Esta representación permite una visualización y un análisis natural de las interacciones, dependencias y estructuras en una amplia gama de dominios.

En la teoría de grafos, un concepto fundamental es el de conjunto independiente, también conocido como conjunto estable, el cual se define como un subconjunto de vértices dentro del grafo en el que no hay dos vértices conectados por una arista. En otras palabras, si se seleccionan dos vértices cualesquiera de un conjunto independiente, no existe una relación o interacción directa entre ellos a través de una arista. Esta propiedad de no adyacencia es central para su aplicación en diversos dominios, (Ballar-Myer, 2019).

Resulta primordial distinguir dos tipos particulares de conjuntos independientes, a saber, un conjunto independiente maximal (MIS, por sus siglas en inglés), definido como un conjunto independiente al que

no se le pueden añadir más vértices sin violar la propiedad de independencia, convirtiéndolo en un conjunto "localmente maximal" y, por otro lado, un conjunto independiente máximo que no es otra cosa que el conjunto independiente maximal con la mayor cardinalidad posible, es decir, aquel con el mayor número de nodos, (Ballar-Myer, 2019; Perdomo, et al, 2015).

Aunque un grafo puede tener múltiples conjuntos independientes maximales, el conjunto independiente máximo es único en tamaño, y su cardinalidad se denomina "número de independencia" del grafo, denotado como $\alpha(G)$, (Ballar-Myer, 2019).

Los conjuntos independientes son fundamentales en teoría de grafos por su capacidad para modelar estructuras sin conflicto, optimizar recursos y analizar relaciones. En el modelado de conflictos, se presentan situaciones donde elementos no deben interactuar, como tareas que no pueden ejecutarse simultáneamente, mientras que, en la optimización, se busca encontrar el conjunto independiente máximo esencial en asignación de recursos, por ejemplo, en redes sociales y en diseño de algoritmos, aunque esto representa un problema NP-difícil, (Academia Lab. 2025).

En el análisis de relaciones existen dos vertientes importantes, agrupamiento (clustering) y clasificación supervisada. En el agrupamiento, se busca agrupar elementos similares sin etiquetas previas, aquí los conjuntos independientes pueden reducir redundancia al seleccionar subconjuntos de datos no conectados (independientes) evitando la duplicación de información. Pueden mejorar la eficiencia al eliminar vértices conectados reduciendo el tamaño del grafo, pero sin perder representatividad. Finalmente, también pueden ser aplicados en el preprocesamiento, técnicas como similarity reduction usan conjuntos independientes para filtrar nodos redundantes antes de aplicar algoritmos como k-means o DBSCAN, (Cela López, 2017).

Por otra parte, en clasificación supervisada, los conjuntos independientes ayudan a seleccionar características relevantes, identificando atributos no correlacionados lo que mejora la precisión del modelo. También permiten diseñar conjuntos de entrenamiento, pues en grafos de relaciones entre instancias, los conjuntos independientes permiten elegir ejemplos representativos sin solapamiento. Por último, ayudan a construir clasificadores eficientes, ya que, en conjuntos de clasificadores heterogéneos, se agrupan por patrones de predicción y se seleccionan subconjuntos independientes para mejorar el rendimiento, (Cela López, 2017).

Enfoque primordial

A pesar de que la TRI ofrece modelos psicométricos más completos, la TCT continúa siendo muy utilizada debido a su sencillez. La TRI involucra modelos logísticos que requieren comprensión de estadística avanzada y para obtener parámetros estables es necesario muestras grandes, algo que no siempre está disponible en entornos educativos pequeños o estudios cualitativos. En cambio, la TCT se basa en fórmulas relativamente sencillas para estimar la puntuación verdadera, la dificultad de los ítems y su índice de discriminación, haciéndola fácil de enseñar y aplicar, especialmente en contextos con recursos limitados.

Otra ventaja de la TCT es que no se requiere software especializado ni modelos matemáticos complejos, esto la hace más accesible para docentes y evaluadores sin formación estadística avanzada. También, los resultados de la TCT son intuitivos para la mayoría de los profesionales en educación, facilitando su interpretación.

El trabajo desarrollado en esta propuesta consiste en combinar la teoría de grafos con modelos psicométricos de la TCT y la TRI, con la finalidad de abrir puertas a nuevas formas innovadoras de analizar evaluaciones educativas, principalmente sobre la calidad de los ítems que conformarán pruebas estandarizadas. La idea principal es explorar una metodología innovadora para la clasificación

de ítems en pruebas estandarizadas, donde se integren parámetros como la proporción de respuestas al ítem ($p - value$), el índice de discriminación (r_{pbis} de la TCT) y el índice de dificultad (b de la TRI), con los conjuntos independientes maximales de la teoría de grafos para la clasificación de los ítems según su dificultad b a partir de los parámetros $p - value$ y r_{pbis} .

Al calcular los MIS del grafo creado utilizando los parámetros psicométricos se pretende identificar subconjuntos de ítems que no compartan similitudes en cuanto a dificultad y discriminación, verificando la concordancia entre la categorización teórica y la empírica. Esto con la finalidad de contar con una herramienta alterna para el diseño y validación de pruebas estandarizadas.

RESULTADOS

Para encontrar los conjuntos independientes máximos (MIS) se utilizó el algoritmo `maximum_independent_set` contenido en `approximation` de la librería `networkx` de Python.

Figura 2

Extracto del algoritmo en Python para obtener los MIS del grafo

```
32 mis_sets = list(nx.algorithms.mis.maximal_independent_set(G))
33 df['es_mis'] = df['item'].apply(lambda x: int(x in mis_sets))
34
35 conjunto_indep = approximation.maximum_independent_set(G)
36 print("Conjunto Independiente Maximal:", conjunto_indep)
```

Al ejecutar el programa se obtuvieron los conjuntos independientes maximales del grafo, tomando únicamente el conjunto independiente máximo cuyos elementos representan ítems con diferencias suficientes para ser considerados distintos tipos de dificultad/discriminación. Los nodos contenidos en el conjunto independiente máximo se muestran a continuación.

{1, 2, 3, 5, 6, 7, 9, 10, 11, 12, 14, 15, 17, 22, 24, 25, 27, 29, 31, 33, 35, 37, 40, 48, 49, 53, 59, 66}

El conjunto independiente máximo encontrado ha sido utilizado para identificar clústeres de cobertura óptima, ya que este conjunto contiene representantes de distintos niveles en dificultad y discriminación, lo que claramente evita la redundancia.

Alta dificultad.

Excelente discriminación. Ítems 29, 31, 37, 40

Buena discriminación. Ítems 3, 10, 11, 14, 24, 31, 59

Regular discriminación. Ítems 5, 6, 12, 15, 25

Dificultad media.

Excelente discriminación. Ítems 2, 9, 27, 33, 35, 49, 66

Buena discriminación. Ítems 7, 17, 22

Baja dificultad.

Baja discriminación. 1, 53

Se puede observar que los clústeres referentes a "Alta dificultad" representan el mayor porcentaje de ítems contenidos en la prueba, sin embargo, en referencia al índice de discriminación, la mayoría de los ítems tienen un nivel bueno o excelente. Esto indica que la prueba construida con estos ítems resulta de un nivel de dificultad alta y el análisis realizado puede ayudar a equilibrar el nivel de dificultad manteniendo la buena distribución de ítems con al menos buena discriminación.

Los resultados anteriores se pueden complementar con las proporciones de aciertos obtenidas por los estudiantes (p –valores de la TCT) como validadores empíricos del nivel de dificultad. Es decir, se puede utilizar el parámetro p –valores como una primera aproximación para el índice de dificultad b proporcionado por la TRI.

Para tal efecto se hace un ajuste en las escalas definidas en la Tabla 1. En este caso se espera que ítems de alta dificultad tengan $p - value < 0.39$.

ítems de dificultad media satisfagan $0.4 \leq p - value \leq 0.7$.

ítems de baja dificultad tengan $p - value > 0.71$.

Esta nueva clasificación permite determinar si los datos empíricos coinciden con la clasificación teórica dada por la TRI. También ayuda a encontrar ítems de mala calidad, ya que si un ítem clasificado como "Alta dificultad" tiene p –value alto, entonces se podrá concluir que está mal calibrado o bien, que está mal diseñado.

Por su puesto, si hay una buena coherencia entre sus categorías y los p –values, se valida la consistencia teórica con evidencia empírica, por lo que pueden usarse como base representativa para diseñar o validar la estructura de una prueba simplemente utilizando el parámetro p –value.

Figura 3

Código para el cálculo del índice de concordancia categoría Teórico/Empírica

```
def etiqueta_p(row):
    if row['p_value'] < 0.39:
        return 'Alta_Dific'
    elif row['p_value'] > 0.7:
        return 'Baja_Dific'
    else:
        return 'Media_Dific'

df['dificultad_empirica'] = df.apply(etiqueta_p, axis = 1)

#Comparar categoría original
coinciden = (df['tipo_item'].values == df['dificultad_empirica'].values).mean()
print(f"Concordancia entre categorías y p_value: {coinciden*100:.1f}%")
```

El índice de concordancia entre la categoría teórica y empírica fue de 89%, el restante 11% está conformado por ítems que presentaron algún problema, entre los que se destacan, doble respuesta correcta y redacción ambigua.

DISCUSIÓN

Los hallazgos confirman que los conjuntos independientes maximales (MIS), construidos desde la teoría de grafos con base en parámetros psicométricos como dificultad (TRI), discriminación (TCT) y

proporción de aciertos (p -value), permiten identificar subconjuntos de ítems suficientemente diversos.

Un conjunto independiente maximal representa un subconjunto de ítems suficientemente diversos entre sí, que no comparten similitudes excesivas en dificultad ni discriminación según los umbrales definidos. Esto los convierte en una herramienta poderosa para el diseño y validación de pruebas.

Al construir un MIS se asegura que ningún ítem del conjunto esté conectado a otro del mismo conjunto, lo que significa la existencia de diferencias suficientes en nivel de dificultad (TRI) y diferencias en discriminación (r_{pbis}).

Esta caracterización coincide con estudios previos que señalan la utilidad del modelado computacional para evitar redundancias en las pruebas (Meijer & Nering, 1999; de la Torre, 2009). La coherencia entre los perfiles teóricos y los datos empíricos sugiere que los MIS son una herramienta válida para representar la diversidad de ítems en una evaluación estandarizada.

En la práctica, un MIS contiene ítems que representan distintos lugares del continuo psicométrico. Por ejemplo, ítems muy fáciles y otros muy difíciles, ítems con discriminación alta y otros con menor discriminación. Si el MIS abarca mayor diversidad que la versión actual de la prueba, podrían estar sobrerrepresentados ciertos tipos de ítems. Si hay ítems fuera del MIS que están altamente correlacionados entre sí, podría haber redundancia innecesaria.

En resumen, el uso de los MIS en el diseño de pruebas presenta varias ventajas, las cuales se detallan a continuación.

Tabla 3

Ventajas del uso de MIS en el diseño de pruebas

Aspecto	Cómo lo aporta el MIS
Cobertura psicométrica	Abarca distintos niveles de dificultad/discriminación.
Reducción de redundancia	Evita incluir ítems con perfiles demasiado similares.
Diversidad cognitiva	Permite inferir distintos niveles de pensamiento/habilidad.
Equilibrio pedagógico	Representa mejor el rango esperado de competencias.

Dentro de las implicaciones Teóricas y Prácticas podemos destacar los siguiente.

Teóricas: Se amplía el marco metodológico de análisis psicométrico al integrar algoritmos de grafos y técnicas de agrupamiento (clustering) como una alternativa de clasificación de ítems. Esto contribuye a la discusión sobre la complementariedad entre TRI y TCT desde una mirada estructural.

Prácticas: Los MIS pueden utilizarse en el rediseño de pruebas, especialmente para seleccionar ítems representativos que cubran distintas regiones del espectro de dificultad y discriminación. Esto favorece el equilibrio pedagógico y la equidad en la medición de aprendizajes.

Dentro de las recomendaciones para el rediseño de pruebas podemos destacar:

- Crear subpruebas temáticas tomando ítems de distintos MIS.
- Construir versiones paralelas usando los ítems que forman diferentes MIS para mantener equivalencia, pero al mismo tiempo la diversidad.
- Evaluar si los MIS coinciden con categorías pedagógicas como “análisis”, “aplicación”, “memorización” o con competencias curriculares.

Dentro de las limitaciones del estudio se encuentran:

El modelo depende de la elección de umbrales de similitud, que pueden variar según la población evaluada y la escala utilizada.

La categorización temática de los ítems requiere una taxonomía coherente y validada por expertos, no siempre disponible, sobre todo si se considera el p -value como primer indicador de dificultad de los ítems.

El problema de encontrar los conjuntos independientes de un grafo es un problema NP-difícil.

Para investigaciones futuras se podría aplicar esta metodología en pruebas reales de alto impacto (por ejemplo, admisión universitaria) para evaluar su capacidad predictiva. También, será necesario comparar los MIS con grupos formados por métodos de agrupamiento (clustering) supervisado o análisis de componentes principales (PCA) para validar su robustez estructural.

Por último, será interesante investigar el impacto de los MIS en la percepción de equidad de los estudiantes y en la eficacia de la retroalimentación formativa.

CONCLUSIÓN

Este trabajo de estudio propone un enfoque interdisciplinario para la clasificación de ítems en pruebas estandarizadas mediante el uso de conjuntos independientes maximales (MIS) extraídos desde estructuras de grafos psicométricos. Los resultados indican que este método permite identificar subconjuntos representativos de ítems con perfiles suficientemente diversos en dificultad, discriminación y proporción de aciertos, lo cual fortalece la cobertura psicométrica y reduce la redundancia evaluativa.

Desde un punto de vista teórico, se integran la Teoría de Respuesta al Ítem (TRI) y la Teoría Clásica del Test (TCT) con herramientas de modelado computacional, abriendo nuevas vías para el análisis estructural de bancos de ítems. Esto se alinea con los avances en evaluación adaptativa descritos por Meijer y Nering (1999), así como con modelos diagnósticos como el DINA, expuesto por de la Torre (2009).

En términos prácticos, los MIS ofrecen una base sólida para el rediseño de instrumentos educativos, facilitando la creación de pruebas balanceadas y adaptadas a distintos niveles de competencia. Esta técnica puede ser especialmente útil en contextos donde la equidad en la medición y la eficiencia diagnóstica son prioridades.

Sin embargo, el estudio presenta limitaciones: la sensibilidad del modelo a los umbrales seleccionados, la necesidad de datos empíricos reales para validación robusta, y la dependencia de categorizaciones temáticas bien definidas.

Es recomendable explorar esta metodología en pruebas de gran escala, comparar su rendimiento con técnicas de agrupamiento supervisadas, e investigar su impacto en la percepción de equidad por parte de los estudiantes.

En suma, la aplicación de conjuntos independientes maximales sobre grafos psicométricos representa una contribución innovadora al campo de la evaluación educativa, al ofrecer una herramienta útil, versátil y coherente con los principios de justicia, precisión y diversidad en el diseño de pruebas.

REFERENCIAS

Abad F, Garrido J, Olea J, Ponsoda V. (2006). Introducción a la psicometría: teoría clásica de los test y teoría de la respuesta al ítem. España: Universidad Autónoma de Madrid.

Academia Lab. (2025). Conjunto independiente (teoría de grafos). Enciclopedia. Revisado el 10 de julio del 2025. <https://academia-lab.com/enciclopedia/conjunto-independiente-teoria-de-grafos>.

AERA/APA/NCME (2014). Standards for Educational and Psychological Testing, Washington, DC: American Educational Research Association/American Psychological Association/National Council on Measurement in Education (AERA/APA/NCME).

Backhoff E, Larrazolo N, Rosas M. (2000) Nivel de dificultad y poder de discriminación del examen de habilidades y conocimientos básicos (EXHCOBA). Revista Electrónica de Investigación Educativa; 2(1), pp. 11 – 29.

Ballar-Myer, J. C. (2019). Deterministic greedy algorithm for maximum independent set problem in graph theory, (2019). <https://www.gcsu.edu/sites/files/page-assets/node-808/attachments/ballardmyer.pdf>

Cela López, J. (2017). Métodos de clustering para conjuntos de clasificación. UAM. Departamento de Ingeniería Informática. <http://hdl.handle.net/10486/680000>.

de la Torre, J. (2009). DINA Model and Parameter Estimation: A Didactic. Journal of Educational and Behavioral Statistics, 34(1), pp. 115 – 130. <https://doi.org/10.3102/1076998607309474>.

Embretson, S. E., & Reise, S. P. (2000). Item response theory for psychologists. Lawrence Erlbaum Associates Publishers.

Haladyna, T. y Rodriguez, M.I (2013). Developing and validating test ítems, Nueva York: Taylor and Francis Group.

Meijer, R. R., & Nering, M. L. (1999). Computerized Adaptive Testing: Overview and Introduction. Applied Psychological Measurement, 23(3), pp. 187 – 194. <https://doi.org/10.1177/01466219922031310>.

Morales P. (2012). Análisis de ítems en las pruebas objetivas. Madrid: Universidad Pontificia Comillas.<http://www.upcomillas.es/personal/peter/otrosdocumentos/analisisitemspruebasobjetivas.pdf>

Moreno, R., Martínez, R. y Muñiz, J. (2004). “Directrices para la construcción de ítems de elección múltiple”, Psicothema, vol. 16, núm. 3, pp. 490 – 497.

Muñiz, J. (2010). Teoría Clásica de los Tests y Teoría de Respuesta a los Ítems. Papeles del Psicólogo, Vol. 31(1), pp. 57 – 66.

Ortiz G, Díaz P, Llanos O, Pérez S, González K. (2015). Dificultad y discriminación de los ítems del examen de Metodología de la Investigación y Estadística. EDUMECENTRO.

Perdomo Flandez, J.A, Bello Lopez, P., Contreras Gonzalez, M., Chávez Benavides, B. (2015). Repositorio de grafos para el conteo de conjuntos independientes. Pistas Educativas, Año XXXVI, No. 112, pp. 1731 – 1750.

van der Linden, W.J. (Ed.). (2016). Handbook of Item Response Theory: Three Volume Set. Chapter 4 Xcalibre. Chapman and Hall/CRC. <https://doi.org/10.1201/9781315119144>.

Weiss, D. J., & Kingsbury, G. G. (1984). Application of computerized adaptive testing to educational problems. *Journal of Educational Measurement*, 21(4), 361–375. <https://doi.org/10.1111/j.1745-3984.1984.tb01040.x>

Todo el contenido de **LATAM Revista Latinoamericana de Ciencias Sociales y Humanidades**, publicados en este sitio está disponibles bajo Licencia Creative Commons .