Universitat d'Alacant
Universidad de Alicante

XARXES D'INVESTIGACIÓ I INNOVACIÓ EN DOCÈNCIA UNIVERSITÀRIA

VOLUM 2025

REDES DE INVESTIGACIÓN E INNOVACIÓN EN DOCENCIA UNIVERSITARIA

VOLUMEN 2025

Satorre Cuerda, Rosana (Coordinación)
Hernández Amorós, María José
Saiz Noeda, Maximiliano
Pellín Buades, Neus (Eds.)

UA | UNIVERSITAT D'ALACANT
UNIVERSIDAD DE ALICANTE
ICE Institut de Ciències de l'Educació
Instituto de Ciencias de la Educación

Redes de Investigación e Innovación en Docencia Universitaria. Volumen 2025

Rosana Satorre Cuerda (Coord.),
María José Hernández Amorós, Maximiliano Saiz Noeda &
Neus Pellín Buades(Eds.)

2025

Redes de Investigación e Innovación en Docencia Universitaria. Volumen 2025

Organització: Institut de Ciències de l'Educació de la Universitat d'Alacant/ *Organización: Instituto de Ciencias de la Educación de la Universidad de Alicante*

Edició / *Edición: Rosana Satorre Cuerda (Coord.), María José Hernández Amorós, Maxiliano Saiz Noeda, Neus Pellin Buades (Eds.)*

Comité tècnic / *Comité técnico*:
Neus Pellin Buades, Universidad de Alicante

Revisió i maquetació: ICE de la Universitat d'Alacant/ Revisión y maquetación: ICE de la Universidad de Alicante
Primera edició: / *Primera edición*: Octubre 2025

ice@ua.es
ISBN: 978-84-09-76835-6

Producció: Institut de Ciències de l'Educació (ICE) de la Universitat d'Alacant / *Producción: Instituto de Ciencias de la Educación (ICE) de la Universidad de Alicante*

20. Evaluación educativa y modelado determinista: Un marco simplificado con la Teoría de Respuesta al Ítem [1]

Rico-Juan, J.R.[1]; Arevalillo-Herráez, M.[2]; Sergio Luján-Mora[1]; Meliá, S.[1]; Navarro Soria, I.[3]

[1]Departamento de Lenguajes y Sistemas Informáticos, Universidad de Alicante
[2] Departamento de Informática, Universidad de Valencia
[3] Departamento de Psicología Evolutiva y Didáctica.

RESUMEN

La evaluación de habilidades en la educación superior mediante tests de respuesta cerrada presenta desafíos significativos debido a la heterogeneidad en la dificultad de los ítems. El objetivo de este trabajo es proponer una metodología determinista que integra elementos de la Teoría de Respuesta al Ítem (IRT) y la Teoría Clásica de los Tests (CTT) para simplificar los cálculos y mejorar la transparencia en la interpretación de los resultados. El método propuesto define la facilidad y dificultad de un ítem a partir de la proporción de respuestas correctas e incorrectas, mientras que la habilidad del participante se establece como la diferencia entre su respuesta y la facilidad del ítem. La capacidad de discriminación de cada ítem se estima mediante la correlación de Pearson entre la habilidad mostrada en el ítem y la habilidad global del estudiante. Esta metodología fue aplicada en dos asignaturas con 88 participantes y un total de 450 ítems. Entre los resultados destacables, el marco propuesto permitió identificar eficazmente los ítems problemáticos y clasificar a los participantes con notable precisión. Además, se logró una reducción significativa del coste computacional y se facilitó la trazabilidad del proceso evaluativo. En conclusión, la metodología se presenta como una alternativa viable que, si bien sacrifica parte de la invarianza de los modelos IRT más complejos, ofrece una solución más sencilla, transparente y computacionalmente eficiente para la evaluación educativa.

PALABRAS CLAVE: Evaluación educativa, Teoría Clásica de los Tests (CTT), Teoría de Respuesta al Ítem (IRT), metodología determinista, calibración de ítems.

1 El presente trabajo ha contado con una ayuda del Programa de Redes de investigación en docencia universitaria del Instituto de Ciencias de la Educación de la Universidad de Alicante (convocatoria 2024). Ref.: 6105

1. INTRODUCCIÓN

Los sistemas basados en test de respuesta cerrada son una herramienta fundamental en la educación superior, utilizados ampliamente para evaluar el conocimiento de los estudiantes en diversas disciplinas (Baker, 2001). Estos sistemas suelen emplear bancos de preguntas clasificadas por temas para generar test personalizados, aplicando una selección aleatoria de ítems (van der Linden & Glas, 2010). Sin embargo, esta estrategia puede introducir sesgos involuntarios, ya que parte del supuesto de que todas las preguntas tienen la misma dificultad. Esto puede repercutir en evaluaciones que no reflejan adecuadamente las habilidades de los estudiantes al tener que contestar ítems con diferente dificultad (Embretson & Reise, 2013).

Para mitigar estos sesgos, una posible solución es calibrar los ítems según su dificultad, permitiendo la creación de tests con una dificultad final homogénea entre los participantes (Hambleton, Swaminathan, & Rogers, 1991). En este contexto, la Teoría de Respuesta al Ítem (IRT, por sus siglas en inglés) se presenta como una herramienta eficaz para la calibración de ítems y la supervisión del proceso de aprendizaje donde se asume cierto error aleatorio en las pruebas y se persigue invarianza en las mediciones (Lord, 2012). La IRT ofrece diversos modelos que permiten estimar parámetros como la dificultad, la discriminación y la habilidad de los participantes, o en nuestro caso, estudiantes (De Ayala, 2013). No obstante, los modelos IRT presentan varios desafíos, como el alto coste computacional, la necesidad de software especializado y la naturaleza no determinista del ajuste de los coeficientes para ítems y participantes (van der Linden, 2017). Estos modelos requieren del uso de técnicas de optimización aproximadas con resultados no deterministas, lo que introduce cierta incertidumbre en los coeficientes calculados, como por ejemplo en el modelo logístico de dos parámetros (2PL) (Baker & Kim, 2004). Además, presentan problemas cuando se consideran ítems con comportamientos de respuesta extremos, como cuando todos los participantes responden correctamente a un ítem o, por el contrario, todos fallan (Samejima, 2008).

Frente a estas limitaciones o problema inherentes al IRT, la Teoría Clásica de los Tests (CTT) ofrece algunas ventajas como un enfoque más sencillo y directo para evaluar la fiabilidad y validez de los test. La CTT se basa en supuestos menos complejos y es más fácil de aplicar en contextos educativos (Crocker & Algina, 1986). Una de las ventajas del CTT es su simplicidad, menor cantidad de cálculos por lo que requiere menor coste computacional, lo cual facilita su implementación y análisis, tanto en entornos con poca cantidad de datos como en los que gestionan grandes cantidades de datos (Thorndike, 1995). Sin embargo, el CTT también tiene limitaciones, como la dependencia de las propiedades de la muestra y la falta de invarianza en los parámetros del ítem (Allen & Yen, 2001).

Para abordar estos desafíos, proponemos una nueva metodología que combina algunas ventajas de ambos enfoques, simplificando las definiciones asociadas a los coeficientes de habilidad, dificultad y discriminación, haciéndolos deterministas, fáciles de calcular e interpretar, además de ser capaz de integrar de forma natural los comportamientos de respuestas extremas. Esta simplificación prioriza la reducción de la incertidumbre en los coeficientes, facilitar la interpretación de los resultados, disminuir el coste computacional y evitar la necesidad de software especializado. Teniendo claro

que se sacrifica la invarianza de los parámetros de modelos IRT.

Las definiciones propuestas para calcular los coeficientes de habilidad de los participantes, y la dificultad y discriminación de los ítems estarán relacionados por fórmulas estadísticas con si la respuestas han sido o no correctas.

Esta metodología no solo simplifica los cálculos y facilita la interpretación, sino que también asegura una trazabilidad clara y una utilidad práctica, especialmente en entornos educativos donde se requiere una evaluación rápida y precisa, tanto en contextos con un número de ítems y participantes reducido como en otros donde puede ser elevado. Además, se pueden calibrar los cuestionarios personalizados para limitar los efectos de excesiva facilidad o dificultad de los ítems asociados a los cuestionarios personalizados.

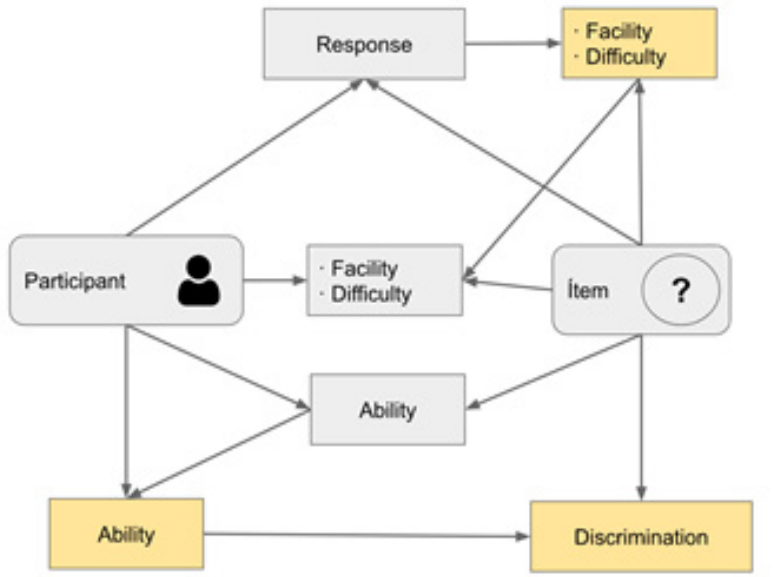
Preguntas de investigación

- RQ1: ¿Sigue teniendo la teoría clásica de los tests (CTT) una relevancia práctica comparada con la teoría de respuesta al ítem (IRT) en las evaluaciones actuales de educación superior?
- RQ2: ¿Cuáles son los beneficios de utilizar la metodología determinista propuesta en comparación con la CTT y la IRT tradicionales en el contexto de cuestionarios de opción múltiple en la educación superior?.

2. MÉTODO

Nuestra propuesta se basa en una simplificación de las hipótesis probabilísticas relacionadas con el IRT. Con ella pretendemos también simplificar los cálculos de los coeficientes asociados, tener una base determinista y evitar los problemas de infraaprendizaje de los modelos cuando se disponen de pocos datos.

Figura 1. Diagram of dependency relationships between the different concepts used in the study. We have highlighted in yellow the main concepts to be calculated.



La Figura 1 muestra las relaciones de dependencia entre los diferentes conceptos involucrados en el IRT. A continuación detallaremos cada uno de ellos y su papel en la metodología.

2.1. DESCRIPCIÓN DEL CONTEXTO Y DE LOS PARTICIPANTES

La metodología se ha aplicado en las asignaturas de Análisis de Datos Clínicos (ADC) en Ingeniería Biomédica a 59 participantes y en Técnicas de Aprendizaje Automático (TAU) del Máster Universitario en Inteligencia Artificial de la Universidad de Alicante a 29 participantes. Inicialmente se intentó usar software especializado en IRT en Python, como py-irt o deep-irt, pero sus implementaciones resultaron limitadas. Por ello, se optó por soluciones en R, encontrando el paquete mirt (Chalmers, 2012), que ofrece modelos 1PL, 2PL, 3PL y otros para cuestionarios con ítems puntuables.

Al implementar el modelo 2PL, surgieron problemas técnicos: ítems con 100% de aciertos provocan errores en los cálculos. Para solucionarlo, se incorporaron dos participantes ficticios — uno denominado bad, que falla todas las respuestas, y otro good, que las acierta — garantizando la existencia de ambos extremos en el cálculo de probabilidades. Esta solución se consideró preferible a eliminar los ítems problemáticos. Sin embargo, el modelo no convergía adecuadamente, y los coeficientes presentaban una gran variabilidad entre entrenamientos.

Tras analizar la relación entre el número de respuestas correctas y la dificultad, se concluyó que plantear un sistema alternativo, más cercano a las respuestas reales, más transparente e interpretable, era la opción más adecuada para este estudio.

Partimos de un banco de ítems categorizados por temas. El número total de ítems disponibles en el banco de preguntas y el número de ítems utilizados por cada test se detallan específicamente en la sección de Resultados (Tablas 1 y 2). A partir de este banco de ítems se generan una serie de cuestionarios tipo test personalizados de respuesta múltiple con 3 opciones donde solo una de ellas es correcta para cada participante (alumno) y atendiendo a una serie de temas. Así que finalmente disponemos de información sobre cada participante y cada ítem como podemos ver en la ecuación 4 donde p es el participante y, i el ítem (se utilizarán estas abreviaturas en adelante).

$$response(p,i) = \begin{cases} 1 \text{ if } p \text{ contesta correctamente a } i \\ 0 \text{ if } p \text{ contesta incorrectamente a } i \\ \varnothing \text{ if } p \text{ no tiene el ítem } i \text{ en el test} \end{cases}$$

A continuación vamos a definir por orden de cálculo las siguientes ecuaciones para calcular los coeficientes relacionados con los ítems y los participantes.

2.2. INSTRUMENTOS Y MODELO DE ANÁLISIS

En esta sección se describen tanto los instrumentos de evaluación utilizados como el modelo determinista propuesto para el análisis de los datos. A continuación, se definen los coeficientes que sirven como métricas para medir la facilidad de los ítems y la habilidad de los participantes. La facilidad y dificultad son complementarios y miden la relación entre respuestas correctas o incorrectas respecto del total (ec. 5 y ec. 6), donde los participantes representan el conjunto con todos los partici-

pantes. El rango de valores de respuestas está comprendido entre 0 y 1.

$$facility(i) = \frac{|response(p,i) = 1 : p\ in\ participants|}{|response(p,i) = 1\ or\ response(p,i) = 0 : p\ in\ participants|} \text{ [ec. 5]}$$

$$difficulty(i) = \frac{|response(p,i) = 0 : p\ in\ participants|}{|response(p,i) = 1\ or\ response(p,i) = 0 : p\ in\ participants|} \text{ [ec. 6]}$$

A partir de estos coeficientes podemos estimar la facilidad o dificultad promedio relacionado con los ítems contestados (correcta o incorrectamente) por un participante (ec. 7 y ec. 8). Estos coeficientes se pueden utilizar para estimar la facilidad o dificultad de los cuestionarios para futuros participantes, y de esta forma corregir desequilibrios, y reducir estas diferencias entre los futuros cuestionarios.

$$facility(p) = \sum facility(i) : i\ in\ items\ of\ p \text{ [ec. 7]}$$

$$difficulty(p) = \sum difficulty(i) : i\ in\ ítems\ of\ p \text{ [ec. 8]}$$

Los coeficientes de habilidad miden por un lado la habilidad del participante en un ítem concreto (ec. 9), y por otro el promedio de esta habilidad para un participante concreto y todos sus ítems contestados (ec. 10) cuyos valores negativos indican menor habilidad y los positivos mayor.

$$ability(p,i) = response(p,i) - facility(i) \text{ [ec. 9]}$$

$$ability(p) = \sum ability(p,i) : i\ en\ ítems\ de\ p \text{ [ec. 10]}$$

Ilustremos la forma de proceder con algunos ejemplos concretos. En los cálculos si un participante ha contestado correctamente un ítem, 1, se le restará la facilidad del mismo (valor entre 0 y 1) y representará lo bien que lo ha hecho respecto al promedio. Ejemplo, un ítem es fácil con un coeficiente de facilidad de 0.9, por lo que su habilidad será de (1 - 0.9) = 0.1, dado que la mayoría lo

han contestado correctamente, por lo que su mérito es escaso, solo de un 0.1; en cambio, si fuera un ítem difícil con una facilidad de 0.2 (solo el 20% lo han contestado correctamente) su habilidad sería de (1 - 0.2) = 0.8, una habilidad alta en este ítem dado que la mayoría de participantes lo han contestado incorrectamente. Continuamos con otro ejemplo, si el participante hubiera fallado el ítem, en el primer caso, con una facilidad de 0.9 su habilidad sería de (0 - 0.9) = -0.9, una habilidad altamente negativa, dado que la mayoría de participantes (90%) han acertado cuando este participante ha fallado en la respuesta a un ítem fácil; y si por el contrario la facilidad del ítem fuera un 0.2 (ítem difícil) y lo fallara, su habilidad sería (0 - 0.2) = -0.2, sería una habilidad negativa, dado que la ha fallado pero escasa, ya que la mayoría de participantes también lo han fallado.

Respecto de la habilidad para un participante se resume su comportamiento acumulado de sus en las habilidades parciales demostradas en cada una de sus respuestas.

Por último, definiremos el coeficiente de discriminación para cada ítem como la correlación (Pearson) existente entre las habilidades de los participantes para cada ítem y respecto de la habilidad general del participante (ec. 11). Los resultados están comprendidos entre -1 y 1, y representan el concepto de coherencia en un ítem de los participantes.

$$discrimination(i) = corr(ability(p), ability(p,i)) : p\ in\ participants \text{ [ec. 11]}$$

De esta forma, si el comportamiento de cada participante de forma individual por ítem está correlacionado con su habilidad acumulada significa que el ítem es adecuado y el resultado tenderá a 1, si por el contrario la habilidad demostrada en un ítem se correlaciona negativamente sobre la general, significa que mayoría de participantes han fallado ese ítem teniendo una buena habilidad o que han acertado ese ítem teniendo una baja habilidad acumulada con lo que el resultado tenderá a -1 y ese ítem debería ser revisado encarecidamente. Un tercer caso es que no hay correlación entre las habilidades parciales y las finales con lo que el ítem también debería ser revisado, ya que no ayuda a discriminar a participantes con habilidades altas o bajas y su comportamiento es similar respecto de dicho ítem.

2.3. PROCEDIMIENTO

El procedimiento de investigación se llevó a cabo siguiendo los siguientes pasos: 1) la administración de los tests en los que se realizaron los cuestionarios tipo test como actividad de teoría en las asignaturas de ADC y TAU a través de la plataforma educativa; 2) recopilación de datos en la que se registraron las respuestas de cada participante para cada ítem de forma dicotómica (correcta=1, incorrecta=0); 3) análisis de datos donde se aplicaron las fórmulas deterministas (ecuaciones 5 a 11) para calcular los coeficientes de facilidad y discriminación de los ítems, así como la habilidad de los participantes; 4) contraste de resultados en la que se contrastaron los coeficientes de habilidad calculados con las calificaciones finales de los estudiantes para observar discrepancias.

Al conocer la facilidad de los ítems una vez contestados los cuestionarios podemos usar estas medidas directamente para calibrarse, o bien, usar un promedio por ítem correspondiente a los

diferentes cursos académicos donde se haya utilizado. De esta forma, podemos diseñar un generador de formularios personalizados con dificultad similar. Una estrategia para ello podría ser generar un número elevado de cuestionarios (mayor al de los participantes), medir y ordenar por su facilidad, y por último seleccionar un rango de las propuestas centrales según el número de participantes con una facilidad similar (descartando los extremos por ser muy fáciles o muy difíciles).

Para este estudio, se utilizaron datos de evaluación académica convenientemente anonimizados, y los resultados son agregados por lo que no es posible la su reidentificación el el anonimato de los participantes está garantizado.

3. RESULTADOS

Los participantes que han usado la plataforma para realizar los test han sido 59 y 29, en las asignaturas de ADC y TAU, respectivamente. Podemos ver un desglose de esta información en las tablas 1 y 2 para cada asignatura. Observamos que los bancos de ítems tenían un total de 272 y 178 respecto de cada una de las asignaturas. Cada participante respondió a un total de 159 ítems distribuidos en 7 cuestionarios, y a un total de 50 ítems distribuidos en 5 cuestionarios en el caso de ADC.

Tabla 1: Distribución de la asignatura Análisis de Datos Clínicos con los temas e ítems asociados disponibles en el banco de preguntas.

Test	Descripción	Ítems disponibles	Ítems por test
1	Introducción a Python	30	20
2	Modelos predictivos	72	30
3	Modelos lineales	30	20
4	Modelos de árboles de decisión	30	20
5	Modelos de conjuntos	50	30
6	Técnicas de validación	40	24
7	Conclusiones finales	20	15
Total		272	159

Tabla 2 : Distribución de temas e ítems asociados a Técnicas de Aprendizaje Automático, incluyendo los disponibles en el banco de preguntas.

Test	Descripción	Ítems disponibles	Ítems por test
1	Introducción a Python	44	10
2	Aprendizaje supervisado (1)	29	10
3	Aprendizaje supervisado (2)	31	10
4	Aprendizaje no supervisado	37	10
5	Aprendizaje por refuerzo	37	10
Total		178	50

En el caso de ADC la tabla 3 muestra el acumulado de habilidad final en orden descendente junto a las calificaciones obtenidas en los cuestionarios. En general, los valores de habilidad y calificación están alineados pero podemos observar cómo existen discrepancias en el orden. Existen numerosos ejemplos marcados con asterisco. Un ejemplo concreto sería en el caso de p17 que con una

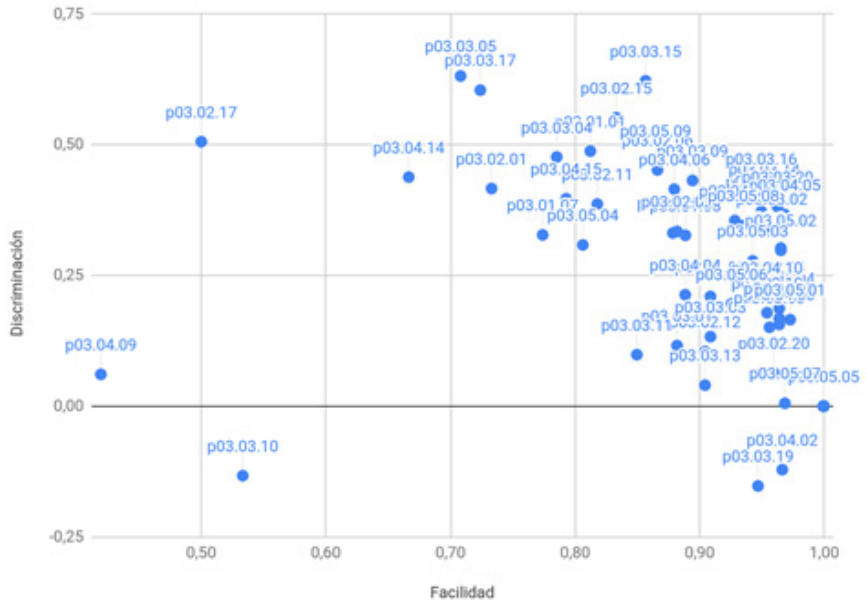
habilidad de 6,65 y una calificación de 9,2, tiene una calificación superior pero una habilidad inferior a p38. Debido a que ha contestado correctamente a ítems más fáciles en los cuestionarios.

Tabla 3 : Habilidades y calificaciones finales obtenidas en los cuestionarios de ADC.Resultados ordenados por habilidades. La discrepancia en el orden está marcado con un asterisco.

Id	Habilidad	Calificación final		Id	Habilidad	Calificación final	
p12	11,15	9,6		p03	1,83	8,6	*
p28	9,55	9,2		p47	1,51	8,8	*
p01	8,40	9,3	*	p34	0,55	8,7	*
p24	8,29	9,4	*	p05	0,48	8,7	*
p54	8,25	9,2		p46	-0,60	8,4	*
p08	7,70	9,1		p25	-1,08	8,6	*
p53	7,13	9,1		p48	-1,45	8,3	*
p50	7,11	9,1		p02	-1,79	8,3	*
p38	6,88	8,9		p44	-1,88	8,2	*
p17	6,65	9,2	*	p45	-2,10	8,3	*
p55	6,34	9,3	*	p29	-2,33	8,3	*
p56	5,47	8,7		p41	-2,46	8,1	*
p21	5,37	8,9	*	p31	-3,24	8,1	*
p32	5,33	8,9	*	p49	-3,63	8,1	*
p19	5,12	8,8	*	p52	-4,19	8,2	*
p13	5,10	9,0	*	p37	-4,34	8,3	*
p43	5,02	9,0	*	p51	-4,98	8,0	*
p27	4,23	8,7		p22	-5,22	7,9	*
p15	3,93	8,8	*	p36	-5,90	8,0	*
p20	3,82	8,8	*	p16	-6,12	8,1	*
p58	3,48	7,8		p26	-6,20	8,0	*
p40	3,46	8,8	*	p09	-6,62	7,9	*
p23	3,30	8,8	*	p04	-6,74	8,0	*
p06	3,06	8,9	*	p42	-6,80	7,8	*
p33	2,57	8,9	*	p30	-8,44	7,7	*
p39	2,41	8,6	*	p14	-11,35	7,8	*
p18	2,09	8,6	*	p57	-15,51	2,6	
p07	1,91	7,3		p11	-15,73	7,1	*
p35	1,90	8,5	*	p10	-28,64	6,4	*

Dado que existen 7 cuestionarios diferentes, vamos a ilustrar un ejemplo del test 3 donde podemos observar qué ítems pueden presentar algún problema de redacción o de referencia a contenidos. La figura 2 muestra un ejemplo donde contrastan la facilidad y la discriminación de forma gráfica. Como se puede observar los ítems con discriminación negativa y menor facilidad nos indicarían que requieren una revisión. Por ejemplo, el ítem etiquetado con p03.03.10 corresponde a la pregunta ¿Qué tipo de modelo se utiliza en lugar del k-nearest neighbors en esta sección? necesitaría una revisión, así como el p03.03.19 y el p03.04.02.

Figura 2: Comparativa entre discriminación y facilidad de los ítems del test 3 de ADC



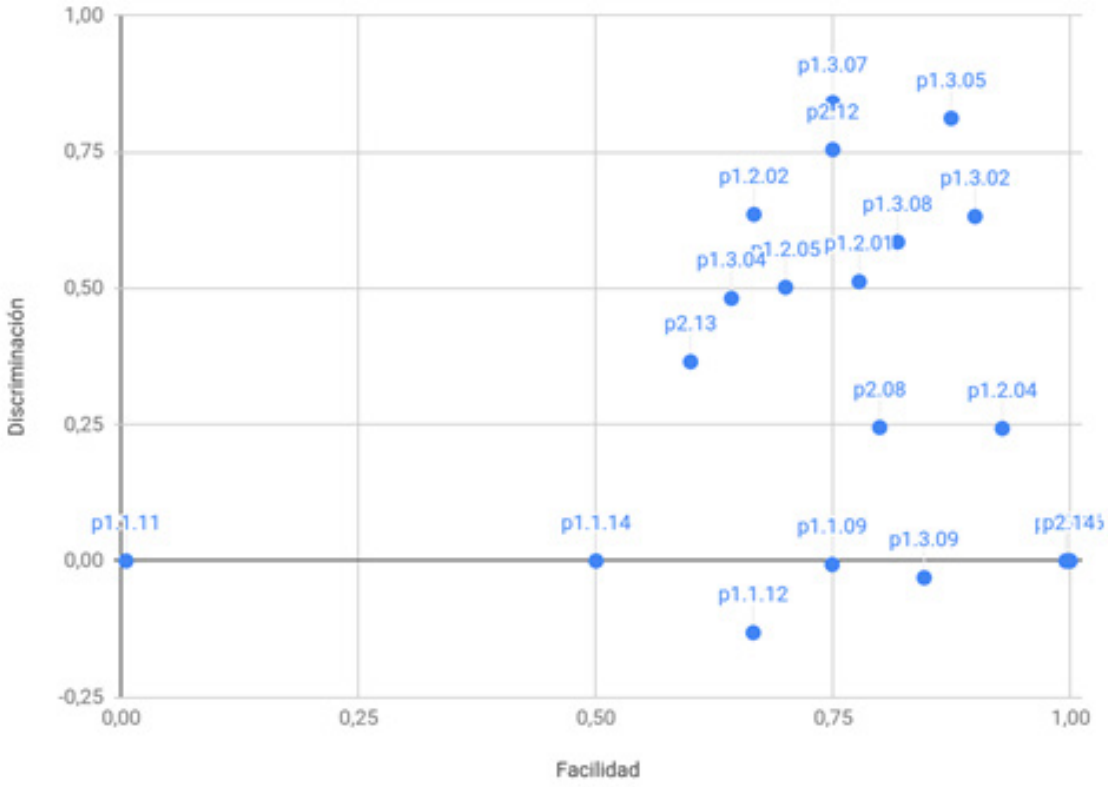
En la asignatura TAU, las habilidades se muestran en la tabla 4, también ordenadas por habilidad en orden descendente (como en el caso de ADC), y también encontramos discrepancias entre los órdenes de habilidad vs. Calificación final. Un ejemplo sería el participante p28 cuya habilidad es muy alta, de 4,29, con una nota de 8,9 que es inferior a la de la p24, p22, p25, p12 y p13 que le preceden con habilidades inferiores.

Tabla 4: Habilidades y calificaciones finales obtenidas en los cuestionarios de TAU.Resultados ordenados por habilidades.

Id	Habilidad	Calificación final		Id	Habilidad	Calificación final	
p03	5,77	10,0		p29	-0,03	8,5	*
p28	4,29	8,9		p23	-0.24	8,6	*
p24	4,12	9,4	*	p17	0,25	7,9	
p22	3,32	9,1	*	p21	0,26	7,9	
p25	2,87	9,7	*	p04	-0,68	8,2	*
p12	2,75	9,5	*	p14	-0,75	8,5	*
p13	2,35	9,1	*	p01	-1,51	8,1	*
p05	2,05	8,9		p19	-1,53	7,7	
p08	1,77	8,8		p10	-1,77	7,9	*
p09	1,47	8,5		p27	-1,97	7,9	*

Al igual que en la asignatura anterior, la figura 3 muestra el contraste entre discriminación y facilidad de los ítems del test 1 en TAU. Prestaremos atención a los ítems con discriminación cercana a cero o negativa, y aquellos que resultan ser menos fáciles. Ej. p1.1.12 que corresponde al ítem El operador lógico habitual que se usa en una condición con dos expresiones que se cumplen al mismo tiempo es:. En este caso parece un exceso de confianza, y la contestación aparentemente parece fácil, no la comprueban con el ordenador y finalmente comenten un error en la respuesta.

Figura 3: Comparativa entre discriminación y facilidad de los ítems del test 1 de TAU



4. DISCUSIÓN Y CONCLUSIONES

El presente estudio demuestra que la Teoría Clásica de los Tests (CTT) sigue siendo una herramienta práctica y relevante (RQ1) en la evaluación educativa actual (Allen & Yen, 2002; Clauser, 2021; DeMars, 2018; Embretson & Reise, 2013; Lance & Vandenberg, 2009). Adicionalmente, y en respuesta a la RQ2, se propone una metodología determinista que, integrando conceptos de la IRT, ofrece una estimación transparente y computacionalmente eficiente de la habilidad, facilidad y discriminación (Chalmers, 2012; van der Linden & Hambleton, 2013). Este enfoque permite al profesorado calcular habilidades de los participantes (alumnado), así como identificar ítems problemáticos y calibrar la dificultad de los tests de siguiente ediciones, promoviendo una evaluación más equitativa (Berg et al., 2017; Schaughency et al., 2012; Shultz, 2020).

Sin embargo, el método presenta limitaciones inherentes. La principal es el sacrificio de la robustez estadística que ofrecen los modelos probabilísticos; al priorizar la simplicidad, se pierde la estimación de la incertidumbre y se corre el riesgo de sobre-simplificar las relaciones en muestras pequeñas (Baker, 2001; Embretson & Reise, 2013; Hambleton et al., 1991; Lord & Novick, 1968). Esto, sumado al tamaño limitado de la muestra del estudio, exige prudencia en la generalización.

En conclusión, esta metodología constituye una alternativa viable y pragmática para evaluaciones personalizadas en educación superior. Se recomienda, como línea de trabajo futuro, explorar la integración de componentes probabilísticos para fortalecer su validez y capacidad de generalización

frente a modelos ya consolidados.

Como trabajos futuros se podría investigar la integración de elementos probabilísticos para mejorar su capacidad de generalización y la validación frente a modelos consolidados.

5. REFERENCIAS

Allen, M. J., & Yen, W. M. (2001). Introduction to measurement theory. Waveland Press.

Baker, F. B. (2001). The Basics of Item Response Theory. ERIC Clearinghouse on Assessment and Evaluation.

Baker, F. B., & Kim, S. H. (2004). Item response theory: Parameter estimation techniques. CRC press.

Berg, D. A., Schaughency, E., van der Meer, J., & Smith, J. K. (2017). Using Classical Test Theory in higher education. In Handbook on measurement, assessment, and evaluation in higher education (pp. 178-190). Routledge.

Birnbaum, Z.W. (1969) On the Importance of Different Components in a Multicomponent System. In: Krishnaiah, P.R., Ed., Multivariate Analysis—II, Academic Press, Waltham, 581-592.

Bock, R. D. (1997). The nominal categories model. In Handbook of modern item response theory (pp. 33-49). New York, NY: Springer New York.

Bond, T. G., & Fox, C. M. (2013). Applying the Rasch model: Fundamental measurement in the human sciences. Psychology Press.

Bradlow, E. T., Wainer, H., & Wang, X. (1999). A Bayesian random effects model for testlets. Psychometrika, 64, 153-168.

Chalmers, R. P. (2012). mirt: A multidimensional item response theory package for the R environment. Journal of statistical Software, 48, 1-29.

Clauser, B. E. (2021). A history of classical test theory. In The History of Educational Measurement (pp. 157-180). Routledge.

Crocker, L., & Algina, J. (1986). Introduction to classical and modern test theory. Holt, Rinehart and Winston, 6277 Sea Harbor Drive, Orlando, FL 32887.

DeMars, C. E. (2018). Classical test theory and item response theory. The Wiley handbook of psychometric testing: A multidisciplinary reference on survey, scale and test development, 49-73.

De Ayala, R. J. (2013). The theory and practice of item response theory. Guilford Publications.

Embretson, S. E., & Reise, S. P. (2013). Item Response Theory for Psychologists. Psychology Press.

Hambleton, R. K., Swaminathan, H., & Rogers, H. J. (1991). Fundamentals of Item Response Theory. Sage Publications.

Hambleton, R. K., & Swaminathan, H. (2013). Item response theory: Principles and applications. Springer Science & Business Media.

Masters, G. N. (1982). A Rasch model for partial credit scoring. Psychometrika, 47(2), 149-174.

Lalor, J. P., & Rodriguez, P. (2023). py-irt: A scalable item response theory library for python. INFORMS Journal on Computing, 35(1), 5-13.

Lance, C. E., & Vandenberg, R. J. (2009). The partial revival of a dead horse? Comparing classical test

theory and item response theory. In Statistical and methodological myths and urban legends (pp. 57-80). Routledge.

Lord, F. M., & Novick, M. R. (1968). Statistical theories of mental test scores. MA: Addison-Wesley.

Lord, F. M. (2012). Applications of item response theory to practical testing problems. Routledge.

Pliakos, K., Joo, S. H., Park, J. Y., Cornillie, F., Vens, C., & Van den Noortgate, W. (2019). Integrating machine learning into item response theory for addressing the cold start problem in adaptive learning systems. Computers & Education, 137, 91-103.

Rasch, G. (1960). Studies in mathematical psychology: I. Probabilistic models for some intelligence and attainment tests.

Samejima, F. (1969). Estimation of latent ability using a response pattern of graded scores. Psychometrika Monograph Supplement, 34(4, Pt. 2), 100.

Samejima, F. (2008). Graded response model based on the logistic positive exponent family of models for dichotomous responses. Psychometrika, 73, 561-578.

Samejima, F. (2011). The general graded response model. In Handbook of polytomous item response theory models (pp. 87-118). Routledge.

Schaughency, E., Smith, J. K., van der Meer, J., & Berg, D. (2012). Classical test theory and higher education: five questions. In Handbook on measurement, assessment, and evaluation in higher education (pp. 137-151). Routledge.

Siegert, R. J., Krägeloh, C. U., & Medvedev, O. N. (2022). Classical Test Theory and the Measurement of Mindfulness. In Handbook of assessment in mindfulness research (pp. 1-14). Cham: Springer International Publishing.

Shultz, K. S., Whitney, D., & Zickar, M. J. (2020). Measurement theory in action: Case studies and exercises. Routledge.

Thorndike, R. M., Cunningham, G. K., Thorndike, R. L., & Hagen, E. P. (1991). Measurement and evaluation in psychology and education. Macmillan Publishing Co, Inc.

van der Linden, W. J., & Glas, C. A. (2010). Elements of adaptive testing (Vol. 10, pp. 978-0). New York, NY: Springer.

van der Linden, W. J. (Ed.). (2017). Handbook of item response theory: Volume 3: Applications. CRC press.

Reise, S. P., & Revicki, D. A. (2014). Handbook of item response theory modeling. New York, NY: Taylor & Francis.

Weiss, D. J., & Kingsbury, G. G. (1984). Application of computerized adaptive testing to educational problems. Journal of educational measurement, 21(4), 361-375.

Weiss, D. J. (2011). Better data from better measurements using computerized adaptive testing. Journal of Methods and Measurement in the Social Sciences, 2(1), 1-27.

Willett, J. B., Singer, J. D., & Martin, N. C. (1998). The design and analysis of longitudinal studies of development and psychopathology in context: Statistical models and methodological recommendations. Development and psychopathology, 10(2), 395-426.

Yeung, C. K. (2019). Deep-IRT: Make deep learning based knowledge tracing explainable using item

response theory. arXiv preprint arXiv:1904.11738.
Zumbo, B. D. (1999). A handbook on the theory and methods of differential item functioning (DIF). Ottawa: National Defense Headquarters, 160.