Análisis comparativo de las Pruebas de Acceso a la Universidad bajo el enfoque de comparabilidad de constructo[1]

Comparative analysis of the University Entrance Examinations within the construct comparability approach

DOI: 10.4438/1988-592X-RE-2020-388-447

Alejandro Veas
Universidad de Alicante
Isabel Benítez
Universidad Loyola Andalucía
Leandro Navas
Raquel Gilar-Corbí
Universidad de Alicante

Resumen

Los procesos de evaluación constituyen una herramienta fundamental en el marco de la formación y selección de estudiantes. Sin embargo, no existen estudios empíricos en España que analicen la utilidad de las pruebas de evaluación para medir el rendimiento académico. El presente estudio, en base a las investigaciones realizadas sobre el enfoque de comparabilidad de constructo (*construct comparability approach*), realiza un análisis comparativo de las calificaciones obtenidas de 15 asignaturas de las Pruebas de Acceso a la Universidad (PAU) en la provincia de Alicante, con una muestra de 6709 estudiantes. Se emplea el modelo de Rasch de crédito parcial como método de estimación, considerando cada materia como un ítem de un instrumento

(1) El presente trabajo se deriva del Proyecto emergente de investigación con referencia GRE17-16, financiado por el Vicerrectorado Investigación y Transferencia de Conocimiento de la Universidad de Alicante.

relacionado con la medición del constructo rendimiento académico. Los resultados iniciales mostraron el cumplimiento de la unidimensionalidad, así como un ajuste de todas materias al modelo, aunque se apreció una falta de discriminación entre sujetos de alto y bajo rendimiento, debido principalmente a la ausencia de monotonocidad de las categorías de puntuación. Se observa que el nivel de dificultad de las materias se adecúa al nivel habilidad de la mayor parte de los sujetos. En base a estos resultados, se destaca la capacidad de las pruebas analizadas para informar sobre el rendimiento académico de los estudiantes. A su vez, se derivan conclusiones relevantes para la mejora de los procesos de calificación, y se proponen investigaciones futuras.

Palabras clave: Pruebas de Acceso a la Universidad, evaluación educativa, rendimiento académico, enfoque de comparabilidad de constructo, modelo de Rasch de crédito parcial.

Abstract

Evaluation processes are a fundamental tool for training and selecting students. However, there are no empirical studies in Spain that analyse the usefulness of assessment tests to measure academic performance. This study, based on research carried out using the construct comparability approach, conducts a comparative analysis of grades achieved by 6709 students pertaining to 15 academic subject areas of the university entrance examinations (*Pruebas de Acceso a la Universidad – PAU*) administered in the province of Alicante regarded as an instrument item related to the measurement of the academic performance construct. The initial results exhibited unidimensionality, and all academic subject areas fit the model, although there was a lack of discrimination between high- and low -performing students, mainly due to the absence of monotonicity in the scoring categories. The difficulty levels of the academic subjects were found to be appropriate for the skill levels of most students. These results demonstrated the ability of the tests analysed to report on the academic performance of the students who took the tests. In addition, important conclusions are presented regarding improvements in the grading processes, and future research studies are proposed.

Keywords: University Entrance Examinations, educational assessment, academic achievement, construct comparability approach, partial credit Rasch model.

Introducción

En los últimos años ha existido un auge creciente en el estudio del rendimiento académico en todos los niveles educativos, ya sea a partir del análisis de variables cognitivas, motivacionales y contextuales involucradas a nivel predictivo o causal (Dicke et al., 2018; Valle et al., 2008), como en el análisis de la calidad de la medición de las pruebas de evaluación externas de rendimiento y sus variables asociadas (Martí y Puertas, 2018; Sayans-Jiménez, Vázquez-Cano, y Bernal-Bravo, 2018). En este último caso, conviene resaltar las mejoras en el estudio del diseño e implementación de pruebas estandarizadas a nivel internacional, tales como TIMMS (*Trends in International Mathematics and Science Study*), PIRLS (*Progress in International Reading Literacy Study*); IALS (*International Assessment of Literacy Survey*), y especialmente PISA (*Programme for International Student Assessment*). Sin embargo, cabe destacar que en España, pese a la profundización en el estudio de diversas variables a partir del análisis de las pruebas mencionadas anteriormente (Elosua, 2013), apenas se han analizado en los últimos años los procesos de medición de calidad de las llamadas Pruebas de Acceso a la Universidad (PAU), más allá de los análisis cuantitativos referidos a diferencias entre grupos de sujetos en ámbitos específicos o de carácter local (Rodríguez-Menéndez, Inda y Peña-Calvo, 2014; Ruiz et al., 2011).

Las PAU constituyen el procedimiento actual para el acceso a estudios universitarios de los estudiantes que obtienen el título de Bachillerato en territorio español, asegurando la selección de alumnos/as en base a una serie de pruebas de aquellas asignaturas troncales y de modalidad que han cursado previamente. En función de la materia que se trate, dichas pruebas tienen diversos formatos, incluyendo comentarios de texto o de imagen, redacciones sobre una temática específica (de extensión corta o larga), o resolución de problemas, entre otros. Además, estas pruebas se diseñan de manera independiente en cada Comunidad Autónoma del país (el alumnado de cada comunidad realiza los mismos modelos de examen), y permite obtener una calificación que, ponderada con la nota del expediente de Bachillerato, se emplea para el cálculo de una nota total utilizada en las posteriores solicitudes de acceso a las distintas titulaciones universitarias; partiendo de la regulación establecida en la Ley Orgánica 8/2013, de 9 de diciembre, para la mejora de la calidad

educativa (LOMCE, 2013), y el Real Decreto 412/2014, de 8 de junio, por el que se establece la normativa básica de los procedimientos de admisión a las enseñanzas universitarias oficiales de grado.

Puesto que las PAU son un proceso de evaluación clave para el futuro de miles de alumnos/as cada año, es imprescindible considerar el papel de la investigación evaluativa en el campo de la educación. En este sentido, partiendo del pragmatismo y del contexto, es necesaria la indagación de los procesos, de los resultados obtenidos, así como de la utilización de los mismos en los planteamientos utilizados por los distintos organismos (Sondergeld y Koskey, 2011) para asegurar los principios de equidad e igualdad de oportunidades en el acceso a los estudios universitarios.

Desde el ámbito de la investigación cuantitativa, se han aplicado diversos métodos estadísticos que permiten indagar sobre el cumplimiento de las condiciones necesarias para asegurar una medición objetiva del rendimiento académico, así como del correcto uso en el diseño de los instrumentos de medida a partir del análisis de las condiciones más específicas. Destacan, por ejemplo, el uso de modelos de valor añadido y modelos multinivel para el análisis del rendimiento académico medido de forma longitudinal (Blanco, González, y Ordóñez, 2009; López-Martín, Kouosmanen, y Gaviria, 2014). Más concretamente, en el ámbito de las PAU, destaca la investigación realizada por Gaviria (2005), en la que analiza mediante diversas técnicas estadísticas (método clásico, método de mínimos cuadrados ordinarios, método multinivel, método de igualación de medias y desviaciones típicas) la equiparación de la nota obtenida del Bachillerato con la obtenida en las PAU, sirviendo esta última como anclaje al ser una prueba común a todo el alumnado. Los resultados muestran que los métodos no clásicos producen mejores resultados que el método de ponderación clásico, aumentando la justicia en la selección del alumnado.

No obstante, pese a la utilidad de ser un conjunto de pruebas comunes para todos los estudiantes que las realizan (teniendo en cuenta la división por ramas de conocimiento), conviene asegurar que las pruebas discriminan adecuadamente el nivel de habilidad de los estudiantes, y que muestran una distribución adecuada de los niveles de dificultad.

En el ámbito de las pruebas de certificación académica, la literatura científica recoge investigaciones relevantes realizadas en otros países que tratan de analizar las propiedades psicométricas descritas previamente, partiendo de diversos modelos teóricos en el análisis de

Recibido: 02-05-2019 Aceptado: 29-11-2019

la comparabilidad de resultados académicos. Destacan especialmente los modelos desarrollados en el Reino Unido; y más concretamente, el enfoque de comparabilidad de rendimiento (Baird, Cresswell, y Newton, 2000), el enfoque de comparabilidad convencional o sociológico (William, 1996b), el enfoque de comparabilidad estadística (William, 1996a), y el elaborado más recientemente que mejora a los anteriores: el enfoque de comparabilidad de constructo (Newton, 2005). Este último modelo indica que, a la hora de comparar dos elementos, sean cuales sean, deben tener algo en común que sirva como base de esta comparación. Al igual que dos tests pueden ser comparados a partir de su medida en una misma escala, en el contexto de la comparación de puntuaciones académicas únicamente podremos comparar aquellas que midan un constructo común; en nuestro caso, el rendimiento académico. Así, la premisa de este enfoque sería la siguiente (Coe, 2008): dos puntuaciones de dos alumnos/as son comparables si el rendimiento académico de ambos, el cual corresponde con el mismo nivel del constructo latente que comparten, da lugar a una misma puntuación. De acuerdo con este postulado, la dificultad de una asignatura corresponderá a un nivel concreto establecido en la variable latente; es decir, una asignatura será más difícil que otra en la medida que, para alcanzar una puntuación concreta, sea necesario un mayor nivel de rendimiento o habilidad (Coe, 2010).

Otros estudios también han mostrado la necesidad de analizar la utilidad de las pruebas de certificación académica diseñadas para la selección de estudiantes y asegurar la comparabilidad de los resultados. Por ejemplo, Hübner, Wagner, Hochweber, Neumann, y Nagengast (2019) mostraron que los resultados recogidos en dos pruebas realizadas a estudiantes en Alemania podían conducir a una selección inadecuada por la falta de ajuste de dichas pruebas a las reformas educativas implementadas. Por otra parte, Korobko, Glas, Bosker y Luyten (2008) encontraron que los resultados obtenidos por estudiantes de los Países Bajos estaban influenciados por las materias elegidas para la evaluación, mostrando la necesidad de ajustar el procedimiento de estimación para evitar una valoración injusta de estudiantes con mayor nivel de rendimiento.

Teniendo en cuenta las investigaciones sobre la temática, la medición de la comparabilidad se basaría en emplear las puntuaciones de las asignaturas como instrumento de medida para la validación del constructo, lo que implica que deben proporcionar buenos niveles de

Recibido: 02-05-2019 Aceptado: 29-11-2019

representatividad de contenido, buena consistencia interna y niveles apropiados de correlación entre el constructo latente y las variables que constituyen las distintas asignaturas.

Este modelo de medición resultaría imposible sin una clara conceptualización del constructo, en nuestro caso el rendimiento académico. Es importante señalar que, pese a ser un concepto ampliamente estudiado, no existe una definición única del mismo en la literatura científica. Dada su complejidad y su enfoque multidisciplinar, la mayoría de las definiciones operativas hacen referencia a la valoración o evaluación de los logros de carácter global obtenidos a nivel escolar (Jiménez, 2000). Nuestro constructo va referido, por tanto, al nivel de logro obtenido a partir del grado de consecución de los estándares de evaluación de las distintas asignaturas de las PAU. Dicho grado de consecución se traduce en unas calificaciones concretas, de tal forma que la comparación del constructo se produce si al aumentar o disminuir la puntuación de una materia supone igualmente avanzar o disminuir en el constructo medido.

El planteamiento teórico presentado en este estudio tiene cabida en el ámbito de la medición que postula el modelo de Rasch (Rasch, 1980; Wright y Stone, 1979), el más conocido dentro de las llamadas Teorías de Respuesta al Ítem (TRI), proporcionando un modelo matemático basado en la calibración de datos ordinales a partir de una escala común de medida, y permitiendo comprobar condiciones tales como la unidimensionalidad, linealidad y monotonocidad. En su forma más básica, este modelo establece que la dificultad de los ítems y la habilidad de los sujetos pueden ser medidas en una misma escala común, y que la probabilidad de que un sujeto acierte un ítem estará condicionada a la diferencia existente entre la habilidad de dicho sujeto y la dificultad de dicho ítem. Ambas medidas (habilidad y dificultad), se examinan en unidades *logit*, puesto que la escala empleada por el modelo es la logarítmica. El uso de una misma escala de medida permite establecer unos intervalos homogéneos, de tal forma que la misma diferencia entre el parámetro de dificultad de un ítem y la habilidad de un sujeto implique la misma probabilidad de éxito a lo largo de toda la escala.

En este nivel de análisis, se parte de la consideración de cada una de las asignaturas como un ítem concreto, siendo el intervalo de puntuación de 0 a 10, lo que implica diversos grados o categorías de éxito. El llamado modelo de crédito parcial (del inglés, *partial credit model*) (Writght y

Recibido: 02-05-2019 Aceptado: 29-11-2019

Masters, 1982), permite analizar la dificultad existente para alcanzar la puntuación específica de cada una de las asignaturas de forma separada siguiendo la metodología Rasch. Dicha metodología se ha empleado en distintas investigaciones en el Reino Unido para el análisis de la comparabilidad de las pruebas de certificación en Educación Secundaria (*General Certificate of Secondary Education*) en estudiantes de 16 años, y las pruebas de certificación de educación avanzada (*General Certificate of Education Advanced*) en estudiantes de 18 años (Coe, 2008; He, Stockford, y Meadows, 2018). La fórmula del modelo es la siguiente:

$$Ln\left(\frac{P_{nij}}{P_{ni(j-1)}}\right) = B_n - D_i F_{ij} = B_n - D_{ij}$$

Siendo:

P_{nij} la probabilidad del sujeto n de acertar el ítem i observado en la categoría j;

B_n la habilidad medida del sujeto n;

D_i la dificultad medida en el ítem i; y

F_{ij} es la calibración medida para el ítem i en la categoría j relativa a la categoría *j-1*, el punto donde las categorías *j-1* y j son igualmente probables en relación con la medida del ítem (Bond y Fox, 2007).

De esta forma, nuestra tarea principal en esta investigación es la aplicación del enfoque de comparabilidad de constructo, desarrollado durante las últimas décadas en el Reino Unido, en las PAU de una provincia perteneciente a una comunidad autónoma española. Concretamente, los objetivos que se persiguen son: 1) comparar los niveles de ajuste y los parámetros de dificultad entre las diversas asignaturas; y 2) comparar la distribución de los niveles de dificultad de las calificaciones de las asignaturas a lo largo del rasgo latente.

Método

Muestra

La muestra está formada por la mayor parte de los estudiantes de la provincia de Alicante que participaron en las PAU en la convocatoria de

Recibido: 02-05-2019 Aceptado: 29-11-2019

junio de 2018. En concreto, se recogieron las calificaciones de un total de 6709 estudiantes examinados en las dos universidades públicas de dicha provincia: la Universidad de Alicante y la Universidad Miguel Hernández de Elche. El porcentaje aproximado de mujeres en ambas universidades oscila en torno al 60 %. Dichas calificaciones se obtienen desde el Servicio de Regulación Universitaria, perteneciente a la Generalitat Valenciana.

Instrumentos

Se emplean las pruebas de las PAU administradas en la provincia de Alicante en la convocatoria de junio de 2018. Dichas pruebas se corresponden con las administradas en el resto de provincias de la comunidad (Valencia y Castellón). Teniendo en cuenta la totalidad de las pruebas correspondientes a las asignaturas troncales y de modalidad, se seleccionan un total de 15 asignaturas, estableciéndose como criterio de selección la inclusión de un mínimo de 600 sujetos por asignatura. Dicho criterio se emplea con el fin de asegurar una mayor precisión de los parámetros estimados (He, Stockford, y Meadows, 2018). De esta forma, las asignaturas empleadas son: biología, castellano: lengua y literatura, cultura audiovisual II, dibujo técnico, economía de la empresa, física, geografía, historia del arte, historia de España, historia de la filosofía, inglés, latín II, matemáticas II, matemáticas aplicadas a las ciencias sociales II, química y valenciano: lengua y literatura.

Las puntuaciones de todas pruebas se establecen en base a unos estándares de corrección previamente establecidos por cada comisión calificadora. De esta forma, existen unos criterios de calificación que se definen a partir de las puntuaciones máximas posibles en cada una de las preguntas formuladas en cada prueba, junto con una instrucción de carácter cualitativa que ayuda a la objetividad de la calificación por parte de los examinadores. Dichos criterios de calificación son públicos y accesibles desde la plataforma web de la Generalitat Valenciana (http://www.ceice.gva.es/va/web/universidad/examenes-y-criterios-de-correccion-de-convocatoria-ordinaria)

Procedimiento

Para el presente estudio, se aplica el enfoque de comparabilidad de constructo, asumiéndose la posibilidad de comparar calificaciones

Recibido: 02-05-2019 Aceptado: 29-11-2019

obtenidas por estudiantes en distintas materias que forman parte de un proceso de evaluación general.

Se emplea el modelo de crédito parcial, usándose el software estadístico Winsteps versión 4.4.0 (Linacre, 2019), cuyas estimaciones se realizan mediante el método de máxima verosimilitud conjunta (Bond, 2004). En dicho modelo, cada una de las asignaturas incluidas se considera un ítem de un mismo instrumento capaz de medir el constructo rendimiento académico.

En primer lugar, de acuerdo con los postulados del modelo de Rasch, se mide la unidimensionalidad del modelo a partir del análisis de componentes principales de las puntuaciones residuales. De acuerdo con Linacre (1998), el valor *(eigenvalue)* obtenido en la comparación de contrastes de residuales no debe ser mayor que 2.

El proceso de estimación de los niveles de dificultad de los ítems (incluyendo sus respectivas categorías) y de habilidad de los sujetos es iterativo, examinando la relación entre la probabilidad de obtener una determinada puntuación en función de la habilidad del estudiante. A partir del procedimiento de máxima verosimilitud es posible obtener el valor, para la dificultad de una determinada puntuación, que mejor explique el patrón de rendimiento registrado. De manera análoga, es posible obtener el valor de habilidad para cada individuo en función del patrón de los índices de dificultad. Este proceso se repite de manera continua empleando las estimaciones de habilidad y dificultad, hasta que la estimación converge.

Mientras que multitud de modelos estadísticos tratan de ajustar el modelo a los datos, en este modelo ocurre lo contrario, es decir, los datos deben ajustarse al modelo para ser aceptados. Este ajuste puede realizarse a partir de las medidas residuales, es decir, de la diferencia entre la respuesta de un sujeto a un determinado ítem y la expectativa de respuesta calculada por el modelo. Las medidas de ajuste pueden ser estandarizadas para un ítem o sujeto concreto de dos formas (Bond y Fox, 2007):

- *Outfit*: es la media cuadrática de los residuales, dividido por los grados de libertad. Esta medida se puede interpretar como una medida global que expresa si las respuestas dadas a un ítem concreto se ajustan al modelo.
- *Infit*: esta medida elimina las puntuaciones extremas que influyen en el *outfit*, de tal forma que emplea los residuales de los individuos

cuyos niveles de habilidad se encuentran en el rango más cercano al ítem concreto.

Los estadísticos *infit* y *outfit* se calculan en base a medias cuadráticas, en función del valor estadístico Chi-cuadrado de Pearson dividido por sus grados de libertad, formando así una escala con valores que pueden oscilar desde 0 a infinito. Valores por debajo de 1 indican un ajuste al modelo superior al esperado, mientras que valores superiores a 1 indican un pobre ajuste del modelo. Así, si tenemos un valor *infit* de 1.40, podemos señalar que hay un 40% más de variabilidad de los datos en comparación con la predicción del modelo; mientras que un *outfit* de .80 indica que existe un 20% menos de variabilidad de los datos observados con respecto a la predicción del modelo.

Se han establecido distintos valores de ajustes en función de los propósitos de análisis (Coe et al., 2008; Tan y Yates, 2007). Linacre (2002) sugirió que los valores superiores a 2 implican forzosamente un mal ajuste del modelo y la imposibilidad de obtener conclusiones fiables del análisis. Por ello, los autores del presente estudio se acogen a este valor de ajuste paramétrico, en ítems y en sujetos; en coincidencia con investigaciones previas bajo el enfoque de comparabilidad de constructo (He, Stockford, y Meadows, 2018). Además, la media de habilidad de los sujetos en las distintas materias se fijó a 0 para que las estimaciones de los parámetros fueran comparables entre ellas.

Resultados

En primer lugar, al comprobar los estadísticos globales del modelo se observó un índice de fiabilidad del sujeto de .74, y un índice de separación del sujeto de 1.69. Estos valores se consideran bajos, e indican que el conjunto de asignaturas no son lo suficientemente sensibles para distinguir de manera eficaz a estudiantes de alto y bajo rendimiento (Bond y Fox, 2007).

Con respecto a la unidimensionalidad del modelo a partir del análisis de componentes principales de las puntuaciones residuales (Bond y Fox, 2007), los resultados muestran un factor principal capaz de explicar el 51.3 % de la varianza del rasgo latente. Con respecto al hipotético segundo factor, muestra un valor inferior a 2 (Eigenvalue V2 = 1.4), lo que confirma la unidimensionalidad del modelo.

En la Tabla I se muestran las asignaturas ordenadas por su parámetro de dificultad (de mayor a menor), así como sus respectivos índices de ajuste. Se observa un ajuste óptimo de todas las materias al modelo, de acuerdo con los criterios establecidos. Las asignaturas con mayor índice de dificultad son, en este orden, Química, seguida de Geografía y Física. Las asignaturas con menor nivel de dificultad son Historia de España, Matemáticas II y Economía.

TABLA I. Parámetros de dificultad y estadísticos de ajuste de las asignaturas de las PAU analizadas

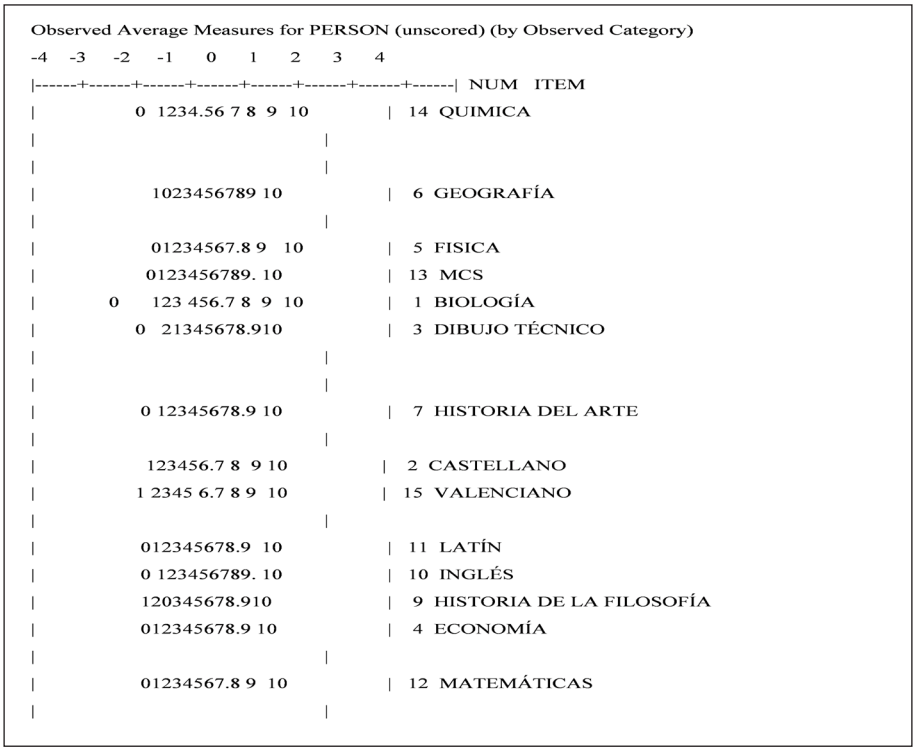
Asignaturas	Nº estudiantes	Dificultad	Infit	Outfit	Correlación ítem-escala
Química	2179	-0.03	0.92	0.91	.76
Geografía	1619	-0.16	1.40	1.40	.62
Física	1504	-0.24	1.12	1.12	.75
Matemáticas aplicadas a las ciencias sociales II	2236	-0.26	1.25	1.25	.63
Biología	1738	-0.29	0.90	0.89	.74
Dibujo Técnico	697	-0.31	1.51	1.50	.62
Historia del Arte	714	-0.41	1.30	1.29	.68
Castellano: Lengua y Literatura II	6123	-0.52	0.71	0.74	.67
Valenciano: Lengua y Literatura II	4747	-0.55	0.69	0.71	.64
Latín	864	-0.63	1.17	1.16	.69
Inglés	5960	-0.65	1.17	1.16	.61
Historia de la Filosofía	778	-0.66	1.37	1.34	.64
Economía	1698	-0.67	0.99	0.98	.70
Matemáticas II	3177	-0.75	1.24	1.19	.68
Historia de España	6124	-0.81	0.90	0.92	.62

Fuente: Elaboración propia basada en los resultados proporcionados por el software Winsteps

Dado que se emplea el modelo de crédito parcial, las calificaciones de todas las materias tienen sus propios índices de ajuste. En este sentido,

todas las calificaciones de cada una de las asignaturas tienen un ajuste óptimo, con valores que oscilan entre 0.7 y 1.9. Sin embargo, al observar la distribución de la dificultad de las calificaciones (Gráfico I), se aprecia que en varias asignaturas no se cumple el criterio de monotonocidad, de forma que al pasar de una calificación a otra inmediatamente superior, no se detecta un aumento en la dificultad asociada. Este hecho se produce en las asignaturas de Historia de España, Historia de la Filosofía, Dibujo técnico y Geografía. Se observa también que en las calificaciones extremas hay una mayor dispersión en la distribución escalar. Por ejemplo, obtener un 10 en Biología es más difícil que obtenerlo en Matemáticas aplicadas a las ciencias sociales. Por otro lado, obtener un 1 en Latín es más difícil que obtenerlo en Valenciano.

GRÁFICO I. Distribución de las calificaciones en el rasgo latente por asignatura

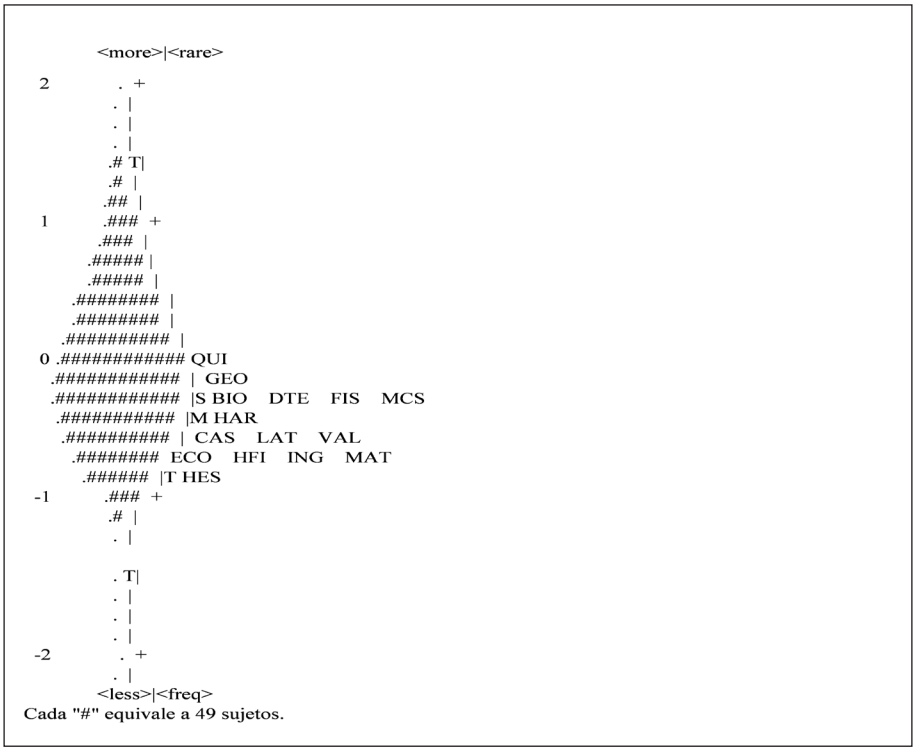


Fuente: Elaboración propia basada en los resultados proporcionados por el software Winsteps

Recibido: 02-05-2019 Aceptado: 29-11-2019

En el Gráfico II se puede visualizar el llamado "mapa de Wright", donde se observa la distribución de sujetos e ítems a lo largo del rango de habilidad y dificultad, respectivamente. Los sujetos están distribuidos en el lado izquierdo del gráfico, mientras que las materias se sitúan en el lado derecho. Se puede observar que la dificultad de las distintas asignaturas se corresponde con el rango de habilidad de los sujetos situado entre los logits 0 y -1. Este hecho es positivo, ya que significa que la mayoría de los sujetos tienen la habilidad suficiente para realizar todos los exámenes. Existe además una proporción de sujetos situados entre los logits 0 y 1, lo que implica que tuvieron un nivel de habilidad mayor, y por tanto más probabilidad de obtener buenas calificaciones en las distintas pruebas.

GRÁFICO II. Mapa de personas e ítems

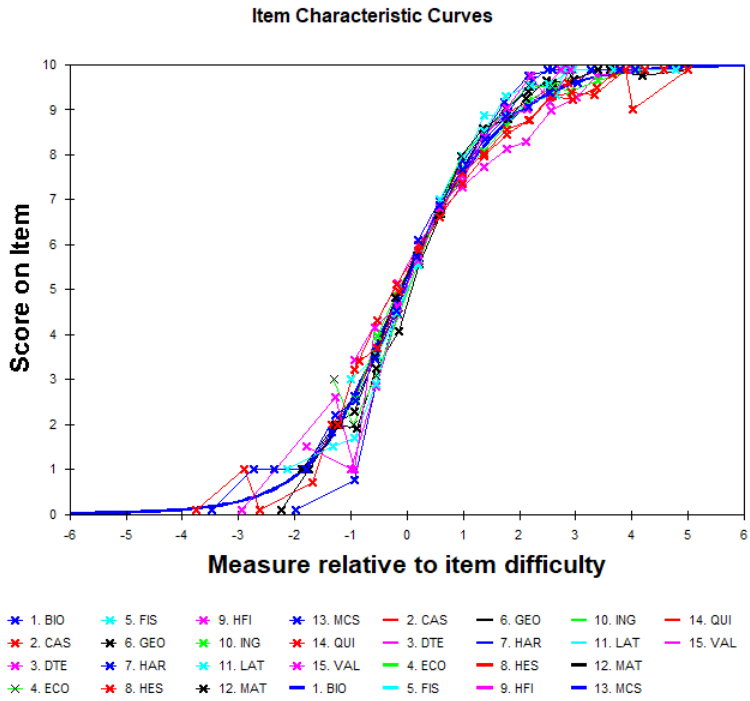


Fuente: Elaboración propia basada en los resultados proporcionados por el software Winsteps

Recibido: 02-05-2019 Aceptado: 29-11-2019

Por último, en el Gráfico III se observan las llamadas Curvas Características de los Ítems (CCI) de todas las materias analizadas. En ellas, se aprecia la relación existente entre las puntuaciones esperadas de cada ítem y las observadas, en función del nivel de habilidad *q* de los sujetos en el rasgo latente. Se aprecia buen ajuste cuando las puntuaciones observadas y esperadas se solapan entre los puntos y la línea continua, respectivamente. En concreto, se detecta un patrón de dificultad similar de las asignaturas en las puntuaciones medias (entre -2 y +2 logits), y diferencias en las puntuaciones extremas (entre -4 y -2 logits y entre +2 y +2 logits). En el lado izquierdo del gráfico se sitúan las asignaturas más fáciles en función del nivel de habilidad requerido; mientras que en el lado derecho se encuentran las puntuaciones que requieren un nivel de habilidad mayor.

GRÁFICO III. Curvas Características de los Ítems



Recibido: 02-05-2019 Aceptado: 29-11-2019

Discusión y conclusiones

Los análisis iniciales muestran el cumplimiento del criterio de unidimensionalidad, el cual es esencial para la aplicación del modelo, y la posibilidad de establecer un constructo latente definido como rendimiento académico, dentro del ámbito de las PAU. Sin embargo, cabe señalar que, pese al establecimiento de este constructo operativo, no se debe interpretar la existencia de un único proceso global. La literatura científica señala que la interpretación de dicho constructo no es clara, ya que no sirve de base para el establecimiento de los propósitos específicos de cada una de las pruebas desarrolladas (He, Stockford y Meadows, 2018). Se puede considerar, por tanto, que todas las pruebas requieren de habilidades específicas, pero al mismo tiempo todas ellas requieren procesos cognitivos globales relacionados con la medición del constructo.

Con respecto al primer objetivo, se observa un ajuste óptimo de todas las materias sometidas a análisis, lo cual permite considerar las propiedades de invarianza asumidas por el modelo de Rasch (Bond y Fox, 2007). Por tanto, las consecuencias de este tipo de estimación residen en la posibilidad de realizar inferencias más allá de la muestra de estudiantes empleada. Al mismo tiempo, el ajuste de las materias permite la comparación en términos de los parámetros de dificultad obtenidos a partir de los niveles de habilidad requeridos para alcanzar cada una de las calificaciones posibles. A partir de estos resultados se deriva una conclusión clave en el ámbito de la evaluación de las PAU, como es la elección de las materias por parte del alumnado; hecho bastante discutido en la literatura internacional (Lamprianou, 2009). Bell et al. (2007) indican que la dificultad percibida por el estudiante hacia una o varias asignaturas puede suponer un obstáculo para el acceso a la universidad. Como consecuencia, otras asignaturas se ven favorecidas con una mayor tasa de matriculación. Teniendo en cuenta los resultados del presente estudio, este hecho podría estar sucediendo con la asignatura de Historia de España en detrimento de Historia de la Filosofía, ya que el alumnado debe elegir una de estas dos materias, y en la primera hay más del triple de candidatos que en la segunda.

El análisis del segundo objetivo resalta la necesidad de considerar la escala de calificaciones empleada en las PAU de manera tradicional, ya que no se cumple con el criterio de monotonocidad típico en las

pruebas de rendimiento. La existencia de 10 categorías no discrimina adecuadamente en puntos concretos del rasgo latente. Cabe destacar que en la mayoría de los países en donde se llevan a cabo análisis comparativos de resultados de pruebas de acceso, emplean un menor número de categorías de calificación. La posibilidad de reajustar el sistema de calificación permitiría a su vez otros posibles cálculos en el mismo marco comparativo, como por ejemplo evaluar la dificultad relativa de cada calificación a partir de la comparación entre ésta y la media de dificultad obtenida de todas las materias, expresando dichas diferencias en la unidad de medida *logit* o en términos de puntuaciones directas. Esta medida permitiría observar la evolución de la dificultad de las puntuaciones de las distintas materias a lo largo de convocatorias sucesivas en distintos cursos académicos (He, Stockford y Meadows, 2018).

En relación con el párrafo anterior, el índice de separación obtenido es bajo, incidiendo en el hecho de que las pruebas no discriminan bien entre sujetos que muestran alto y bajo nivel del rasgo latente. Sin embargo, el mapa de "Wright" indica que los niveles de dificultad de todas las pruebas se encuentran dentro del rango de habilidad de los sujetos, por lo que existen niveles adecuados de probabilidad de obtener resultados positivos. La situación de las materias en la escala se corresponde con una distribución similar de las categorías en el constructo latente, tal y como se aprecia en las CCI, aunque cabe señalar ciertas diferencias en la distribución de las categorías extremas. De nuevo, estos resultados inciden en la necesidad de recodificar el sistema de categorías para una mejora de la discriminación, al poder incluir un mayor número de estudiantes en cada una de las categorías de bajo y alto rendimiento.

Como conclusión, el presente trabajo trata de iniciar en España un análisis efectivo de comparación de las calificaciones bajo el enfoque de comparabilidad de constructo llevado a cabo en otros países. Conviene tener en cuenta, no obstante, algunas limitaciones que pueden orientar dicha temática a futuras investigaciones. En primer lugar, cabe destacar que las muestras empleadas en otros países son mucho más amplias, al recoger datos de carácter nacional; lo cual permite obtener una mejor estimación, dado el mayor número de calificaciones y de materias. La realización del presente estudio a escala provincial permite un acercamiento al análisis de la comparabilidad de constructo, confirmando la posibilidad de realizar estudios futuros análogos a los

Recibido: 02-05-2019 Aceptado: 29-11-2019

realizados en otros países; y en nuestro contexto específico, estableciendo comparaciones entre comunidades autónomas para poder establecer las medidas de equidad oportunas. En este sentido, es necesario analizar la influencia de diversos factores diferenciales, tales como la selección individual de las materias o el efecto de las reformas educativas en la evaluación (Hübner et al., 2019; Korobko, Glas, y Bosker, 2008). Por otro lado, no se ha tenido en cuenta la posible influencia de los evaluadores en la medición del modelo. Esta posibilidad no se ha explorado aún en la literatura científica dentro de este ámbito. Sin embargo, teniendo en cuenta que la mayoría de las PAU son de redacción; las diferencias entre los calificadores en la interpretación de las tareas y de las categorías de evaluación, así como otros posibles efectos (efecto halo, sesgo de género y cultura, etc.), pueden contribuir al error de medida, a la validez y a la justicia de las evaluaciones (Prieto, 2011). El modelo de Rasch permite su consideración a partir de una extensión del modelo de crédito parcial, denominado modelo de Rasch de múltiples facetas (del inglés, *Many Facet Rasch Measurement*).

Referencias bibliográficas

Baird, J., Cresswell, M., y Newton, P. (2000). Would the real gold standard please step forward? *Research Papers in Education, 15*(2), 213-229. https://doi.org/10.1080/026715200402506

Blanco, A., González, C., y Ordóñez, X. G. (2009). Patrones de correlación entre medidas de rendimiento escolar en evaluaciones longitudinales: un estudio de simulación desde un enfoque multinivel. *Revista de Educación, 348*, 195-215.

Bond, T. (2004). Validity and assessment: A Rasch measurement perspective. *Metodología de las Ciencias del Comportamiento, 5*, 179-194.

Bond, T. G., y Fox, C. M. (2007). *Applying the Rasch model: Fundamental measurement in the human sciences.* New York, USA: Psychology Press.

Coe, R. (2008). Comparability of GCSE examination in different subjects: an application of the Rasch model. *Oxford Review of Education, 24*(55), 609-636. https://doi.org/10.1080/03054980801970312

Coe, R. (2010). Understanding comparability of examination standards. *Research Papers in Education, 25*(3), 271-284. https://doi.org/10.1080/02671522.2010.498143

Dicke, T., Marsh, H. W., Parker, P. D., Pekrun, R., Guo, J., y Televantou, I. (2018). Effects of school-average achievement on individual self-concept and achievement: Unmasking phantom effects masquerading as true compositional effects. *Journal of Educational Psychology, 110*(8), 1112-1126. https://doi.org/10.1037/edu0000259

Gaviria, J. L. (2005). La equiparación del expediente de Bachillerato en el proceso de selección de alumnos para el acceso a la universidad. *Revista de Educación, 337*, 351-387.

He, Q., Stockford, I., y Meadows, M. (2018). Inter-subject comparability of examination standards in GCSE and GCE in England. *Oxford Review of Education, 44*(4), 494-513. https://doi.org/10.1080/03054985.2018.1430562

Hübner, N., Wagner, W., Hochweber, J., Neumann, M., y Nagengast, B. (2019). Comparing apples and oranges: Curricular intensification reforms can change the meaning of students' grades! *Journal of Educational Psychology.* https://doi.org/10.1037/edu0000351

Jiménez, M. (2000). Competencia social: intervención preventiva en la escuela. *Revista Infancia y Sociedad, 24*, 21-48.

Korobko, O. B., Glas, C. A., Bosker, R. J., & Luyten, J. W. (2008). Comparing the difficulty of examination subjects with item response theory. *Journal of Educational Measurement*, *45*(2), 139-157.

Lamprianou, I. (2009). Comparability of examination standards between subjects: an international perspective. *Oxford Review of Education, 35*(2), 20-226. https://doi.org/10.1080/02054980802649360

Ley Orgánica 8/2013, de 9 de diciembre, para la Mejora de la Calidad Educativa. *Boletín Oficial del Estado (España)*, 10 de diciembre de 2013, 295, 97.858-97.921.

Linacre, J. M. (1998). Structure in Rasch residuals: Why principal component analysis? *Rasch Measurement Transactions, 12,* 636.

Linacre, J. M. (2019). *WINSTEPS rasch measurement computer program* [version 4.4.0]. Chicago, USA: Winsteps.

López-Martín, E., Kuosmanen, T., y Gaviria, J. L. (2014). Linear and nonlinear growth models for value-added assessment. An application to Spanish primary and secondary schools' progress in reading comprehension. *Educational Assessment, Evaluation and Accountability, 26*(4), 361-391. https://doi.org/10.1007/s11092-014-9194-1

Martí, M. L., y Puertas, R. (2018). Comparativa de la eficiencia educativa de Europa y Asia: TIMMS 2015. *Revista de Educación, 380*, 45-74. https://doi.org10.4438/1988-592X-RE-2017-380-372

Newton, P. E. (2005). Examination standards and the limits of linking. *Assessment in Education, 12*(2), 105-123. https://doi.org/10.1080/09695940500143795

Rasch, G. (1980). *Probabilistic models for intelligence and attainment tests* (Copenhagen, Danish Institute for Educational Research). Expanded edition (1989) with foreword and afterword by B. D. Wright. Chicago, USA: The University of Chicago Press.

Real Decreto 1105/2014, de 26 de diciembre, por el que se establece el currículo básico de la Educación Secundaria Obligatoria y del Bachillerato. *Boletín Oficial del Estado (España)*, 3 de enero de 2015, 3, 169-546.

Real Decreto 412/2014, de 8 de junio, por el que se establece la normativa básica de los procedimientos de admisión a las enseñanzas universitarias oficiales de grado. *Boletín Oficial del Estado (España)*, 7 de junio de 2014, 138, 43307-43323.

Rodríguez-Menéndez, M. C., Inda, M. M., y Peña-Calvo, J. V. (2014). Rendimiento en la PAU y elección de estudios científico-tecnológicos en razón de género. *Revista Española de Orientación y Psicopedagogía, 25*(1), 111-127.

Ruiz, J., Dávila, P., Etxeberría, J., y Sarasua, J. (2011). Pruebas de selectividad en matemáticas en la UPC-EHU. Resultados y opiniones de los profesores. *Revista de Educación, 362*, 217-246.

Sayans-Jiménez, P., Vázquez-Cano, E., y Bernal-Bravo, C. (2018). Influencia de la riqueza familiar en el rendimiento lector del alumnado en PISA. *Revista de Educación, 380*, 129-155. https://doi.org/10.4438/1988-592X-RE-2017-380-375

Sondergeld, T., y Koskey, K. (2011). Evaluating the impact f an urban comprehensive school reform: An illustration of the need for mixed methods. *Studies in Educational Evaluation, 37*, 91-107. https://doi.org/10.1016/j.stueduc.2011.08.001

Tasmanian Qualifications Authority (2007). *How the scaled awards are calculated and used to determine the tertiary entrance score.* Available online at: www.tqa.tas.gov.au/0477

Tognolini, J., y Andrich, D. (1996). Analysis of profiles of students applying for entrance to universities. *Applied Measurement in Education, 9*(4), 323-353. https://doi.org/10.1207/s1532481ame0904_3

Valle, A., Núñez, J. C., Cabanach, R. G., González-Pienda, J. A., Rodríguez, S., Rosário, P., … Muñoz-Cadavid, M. (2008). Self-regulated profiles and academic achievement. *Psicothema, 20*(4), 724-731.

Wiliam, D. (1996a). Meanings and consequences in standard setting. *Assessment in Education, 3*(3), 287-308. https://doi.org/10.1080(0969594960030303

Wiliam, D. (1996b). Standards in examinations: A matter of trust? *The Curriculum Journal, 7*(3), 293-306.

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

Información de contacto: Alejandro Veas Iniesta, Universidad de Alicante, Facultad de educación, departamento de psicología evolutiva y didáctica, Carretera San Vicente del Raspeig, s/n, CP: 03690 E-mail: alejandro.veas@ua.es

Recibido: 02-05-2019 Aceptado: 29-11-2019