Validación interna de la prueba de inglés de la EvAU en la Comunidad de Madrid

Miguel Fernández Álvarez
Universidad Politécnica de Madrid

Judit Ruiz-Lázaro
Universidad Europea de Madrid

Jesús García Laborda
Universidad de Alcalá

Received: 2022-12-26 / Accepted: 2023-09-10
DOI: *https://doi.org/10.30827/portalin.vi41.26977*
ISSN paper edition: 1697-7467, ISSN digital edition: 2695-8244

RESUMEN: La evaluación de la competencia lingüística para acceder a la universidad en España se modificó en el año 2009 y, actualmente, se trata de un tema social y académico de interés debido a las continuas propuestas para sus características y su diseño. El presente estudio tiene el objetivo de analizar la validez interna de la prueba de Lengua Extranjera-Inglés de la Evaluación para el Acceso a la Universidad que se ha diseñado y se ha aplicado en la Comunidad de Madrid en el año 2021. Para ello, se plantea un diseño no experimental y transversal de carácter descriptivo. Se aplicó la prueba de Lengua Extranjera-Inglés a 4,243 aspirantes a los estudios superiores universitarios en el mes de junio de 2021, de los cuales 3,746 resultaron válidos para el presente estudio. Se llevaron a cabo análisis estadísticos descriptivos, análisis de la fiabilidad y el empleo del Modelo de Rasch para la comparación de puntuaciones obtenidas por los aspirantes en la prueba a través de un Mapa de Wright que muestra la distribución de los niveles de dificultad de los ítems. Los resultados aportan evidencias sobre la calidad técnica de la prueba de Lengua Extranjera-Inglés. Se observa que las preguntas de verdadero/falso podrían estar mostrando dificultades técnicas por ser los ítems más sencillos, así como los de la localización de vocabulario que, por el contrario, resultarían ser los más complejos. Asimismo, a pesar de que la fiabilidad es adecuada, la disparidad de resultados en las preguntas que demandan destrezas en la expresión escrita sugiere que deben plantearse instrumentos de evaluación analíticos y objetivos. Los resultados sugieren la necesidad de un replanteamiento de los ítems, así como la evaluación de todas las destrezas lingüísticas para dar respuesta al Marco Común Europeo de Referencia para las Lenguas.
Palabras clave: validez, pruebas externas, acceso a la universidad, lengua extranjera, lengua inglesa

Internal validation of the EvAU English test in the Comunidad de Madrid

ABSTRACT: The assessment of linguistic competence for access to university in Spain was modified in 2009 and is currently a social and academic topic of interest, as proposals are constantly being made regarding its characteristics and design. The aim of this study is to analyze the internal validity of the foreign language proficiency assessment test for university entrance in the Madrid region in 2021. It is a non-experimental cross-sectional study with descriptive character. The foreign language test for English was used with 4,243

university applicants in June 2021, of which 3,746 were valid for the present study. Descriptive statistical analyses, reliability analyses, and the application of the Rasch model were performed to compare the scores obtained by the applicants in the test using a Wright map showing the distribution of the difficulty levels of the items. The results provide information about the technical quality of the English foreign language test. It can be seen that the true/false questions could pose technical difficulties since they are the easiest tasks, as well as the vocabulary questions, which, in contrast, would prove to be the most complex. Also, the disparity in results for questions requiring written expression suggests that analytical and objective assessment tools should be considered, even if reliability is adequate. The results suggest that both the tasks and the assessment of all language skills need to be reconsidered in order to comply with the Common European Framework of Reference for Languages.
Keywords: validity, external testing, access to university, foreign language, English language

1. Introducción

Vivimos unos tiempos complejos en los que la prueba de acceso a la universidad o prueba final de Bachillerato están siendo cuestionadas por su ínfima capacidad de validez predictiva, los indefinidos de sus objetivos o la carente validez, entre otros muchos aspectos que deben estar presentes en cualquier examen de lenguas, sobre todo en aquellos considerados de alto impacto, como es el examen que determina el acceso a la universidad. Son numerosos los intentos que ha habido en los últimos años de revisar esta prueba. Algunos problemas a los que se ha enfrentado su diseño están relacionados, principalmente, con la falta de investigación de la misma. Algunos estudios se centran en el diseño (Fernández Álvarez, 2007; García Laborda & Martín-Monje, 2013; Ruiz-Lázaro et al., 2021), en la introducción de la tecnología (García Laborda & Gimeno Sanz, 2008; García Laborda et al., 2011), en las novedades técnicas (Sevilla-Pavón et al., 2017) o en propuestas competenciales (Amengual-Pizarro & Méndez García, 2012; Bueno & Luque, 2012; Díez Bedmar, 2011), entre otros. Sin embargo, se desconocen estudios previos sobre el análisis de la validez interna de la prueba.

Una de las reformas más destacadas es la propuesta por la Conferencia de Rectores de las Universidades Españolas (CRUE, 2022) debido a los cambios que supone la inclusión del denominado "ejercicio de madurez". Asimismo, destaca la restructuración de los ejercicios que componen cada uno de los exámenes de esta prueba, siendo uno de los principales cambios la inclusión de preguntas en distintos idiomas con el fin de evaluar las competencias lingüísticas.

Independientemente de los cambios que se propongan y/o se hagan al examen de lengua extranjera (y aquí nos referiremos al de inglés), este debe estar regido por todos los principios básicos que deben regir cualquier examen, tales como la validez (Alderson et al., 1995; Bachman, 1990; Bachman & Palmer, 1996), la fiabilidad (Fulcher, 1997; Hughes, 2003) o la viabilidad (Brown, 1994), por mencionar solo aquellos más relevantes. Se trata, por tanto, de medidas mínimas que debe tener una prueba con una influencia decisiva en la vida de una parte importante de la juventud en nuestro país.

En respuesta a la falta de estudios centrados en la validez del examen de inglés de las pruebas de acceso a la universidad, tal y como se ha mencionado anteriormente, este artículo se centra en un estudio de validación interna de la prueba de inglés de la Evaluación

para el Acceso a la Universidad (en adelante, EvAU) diseñado y aplicado en la Comunidad de Madrid. Este estudio servirá para complementar aquellos ya existentes y puede servir como base que ayude a determinar los cambios que se hagan a la nueva prueba de acceso a la universidad.

2. Las pruebas de acceso a la universidad en España

La evaluación para el acceso a la universidad se implantó en el curso académico 2016-2017 como un conjunto de pruebas que deben superarse para acceder a los estudios superiores universitarios en España (Real Decreto-ley 5/2016). Las tres materias comunes que todos los estudiantes deben realizar, independientemente de la comunidad autónoma en la que se presenten, son: (1) Lengua Castellana y Literatura, (2) Lengua Extranjera e (3) Historia de España. Se entiende por lengua extranjera aquella lengua cursada durante el Bachillerato distinta al castellano: alemán, francés, inglés, italiano o portugués. Sin embargo, cabe señalar que, según los datos que publica de forma anual el Sistema Integrado de Información Universitaria, más del 97% de los examinandos elijen Inglés como lengua extranjera (Sistema Integrado de Información Universitaria, 2022).

Cada comunidad autónoma tiene autonomía para diseñar sus pruebas siguiendo una serie de directrices generales (Real Decreto 310/2016), basadas en las enseñanzas mínimas del Bachillerato recogidas en el Real Decreto 243/2022. La prueba de inglés se enmarca dentro del complejo de otros exámenes (Figura 1) que conforman un conjunto y darán como resultado una única nota.

Figura 1. *Estructura de las pruebas de acceso a la Universidad (EvAU)*
Fuente: https://www.eusa.es/blog/claves-para-la-selectividad/

2.1. El examen de inglés en la EvAU

La prueba de inglés de la EvAU está diseñada para identificar las fortalezas y debilidades de los estudiantes y apoyar la toma de decisiones sobre la elegibilidad de aquellos más preparados para el acceso a la universidad. Como herramienta de evaluación sumativa de lengua, la prueba general tiene como objetivo proporcionar información oportuna para medir

la competencia de dicha lengua en situaciones sociales y en algunas ocasiones, académicas. De esta manera serviría para conocer si se han cumplido los objetivos de la asignatura en Bachillerato. Idealmente, como IB TOEFL, debería servir de indicador de si el estudiante será capaz de usar esa lengua adecuadamente en sus estudios universitarios. La sección de inglés, y por extensión las demás, se basa teóricamente en el procedimiento denominado *school-based decision-making* (Carr-Hill et al., 2018), cuyo enfoque es el análisis y propuestas de expertos. Desafortunadamente, en general, los expertos carecen de los conocimientos técnicos y académicos en evaluación para la responsabilidad que supone un examen con un impacto como este (Veas et al., 2020b). Además, estos expertos coordinan las correcciones, de modo que deben basarse en ítems que tienen muy poca variación (objetivos) pero que, al ser mayormente ítems abiertos, introducen sesgos muy importantes que investigadoras como Amengual Pizarro (2004) han probado que ponen en riesgo el grado de fiabilidad de los exámenes (Amengual Pizarro, 2005).

Un problema añadido es que no se generan datos de investigación ni nuevos tipos de ítems desde los años 90 por lo que no hay triangulación de datos ni estudios del funcionamiento de ítems como el que esta investigación trata de hacer. Adicionalmente, tampoco existe una serie de modelos de calificación avanzados, y la ya citada subjetividad entre correctores es un mal significativo, especialmente porque muchos asumen calificaciones holísticas en vez de analíticas (Amengual Pizarro, 2006; Ruiz-Lázaro & González Barbera, 2017).

En la Comunidad de Madrid, para respuestas que implican el desarrollo de la expresión escrita, se utilizan varios modelos de puntuación: el ítem 2 utiliza un sistema holístico mientras que el 5 (redacción de 150-200 palabras) cuenta con una rúbrica genérica, tal y como se observa en la Figura 2.

	Excelente	Nota	Deficiente
CONTENIDO	El mensaje es claro, preciso y coherente, con ideas interesantes, que se atienen al tema psropuesto. Se sigue el requisito de extensión mínima.	---/0,5	**El mensaje es demasiado confuso, ambiguo o incoherente, con ideas irrel vantes o repetitivas. No se sigue el requisito de extensión mínima.**
	Se muestra capacidad para desarrollar un punto de vista personal, con opiniones originales. Las ideas se ilustran de forma adecuada.	---/0,5	**Es difícil distinguir la postura personal del autor. Se incluyen generalidades sin fundamento, porque no se aportan datos o ejemplos que ilustren las ideas expuestas.**
FORMA	Se emplean conectores de forma efectiva y variada.	---/0,5	**Faltan conectores adecuados y se acusa una falta de transiciones temáticas lógicas.**
	No hay errores importantes de gramática.	---/0,5	**Hay errores graves de gramática.**
	No muestra limitaciones en el uso delvocabulario que utiliza.	---/0,5	**Hay errores graves de léxico.**
	No hay errores importantes de ortografía y/o puntación.	---/0,5	**Hay múltiples equivocaciones en el uso de laortografía y/o la puntuación.**
Total		**---/3**	

Figura 2. *Rúbrica del ítem 5*
Fuente: Coordinación de la EvAU de la Comunidad de Madrid

Por otro lado, las reuniones de afinamiento y unificación de criterios de corrección vienen a durar apenas unos minutos (generalmente menos de una hora al final del primer día de exámenes), cuando los vocales de los tribunales han estado vigilando exámenes todo el día. Respecto al seguimiento y las pruebas de seguridad, no se da más que la del traslado seguro y la destrucción de los distintos modelos de cada examen. Como se ha mencionado anteriormente, el examen no se calibra con profesionalidad y ni tan siquiera se pilota. Su diseño se basa simplemente en la experiencia de los comités de los exámenes. Tampoco se realizan evaluaciones de la fiabilidad de los ítems, y hasta la fecha solamente se ha usado el de papel-bolígrafo, a pesar de las numerosas propuestas de modernización de varios autores (García Laborda, 2010; Fernández Álvarez et al., 2022).

2.2. La validez del examen de inglés en la EvAU

El constructo de la prueba de Inglés de la EvAU no está definido en profundidad, incluyendo los descriptores y objetivos de cada ítem (qué quieren medir y cómo). Que la prueba tenga un nivel B1 parece un objetivo no demostrado, ya que, como se mencionó anteriormente, en el diseño de la prueba no trabajan auténticos especialistas profesionales en evaluación y prevalecen las opiniones subjetivas de los miembros del comité de expertos. No existe tampoco un manual técnico que presente los dominios de uso del lenguaje críticos para el desempeño adecuado de la lengua extranjera ni en el uso social ni en el interpersonal o el académico. En la Comunidad de Madrid no hay siquiera una base de datos de ítems sobre la que trabajar de un año a otro. Los pocos estudios que hay sobre la EvAU tampoco muestran Que se haya investigado hasta qué punto las inferencias extraídas serían contrastables en una prueba de validez externa o alineación con otros exámenes del mismo nivel; por ejemplo, a través de la corrección de la sección escrita de la EvAU con correctores del examen *B1 Preliminary*, antes denominado *Preliminary English Test (PET*). Asimismo, tampoco se han estudiado las características específicas de los textos de lectura en aspectos como formas, funciones y arreglos pragmáticos del lenguaje. (p. Ej., finalidad comunicativa, iniciación-respuesta, estructura de participación y registro).

Por supuesto, al no haber pruebas de comprensión o expresión oral (salvo en el caso de Galicia y Cataluña), el uso del lenguaje en las interacciones entre pares y grupos no se aborda de manera adecuada en el aula en el curso de segundo de Bachillerato ("efecto rebote negativo" o *negative washback*) (García Laborda, 2013). Debido a que los estudiantes tienen una evaluación muy parcial, dado que las destrezas practicadas en el aula son generalmente incompletas y que las citadas funciones también son necesarias para que los estudiantes sigan adecuadamente sus cursos y participen activamente en las interacciones académicas y sociales en su periodo universitario, muchos alumnos se ven claramente discriminados por un sistema que "no reconoce" su competencia real en lengua extranjera. La cobertura insuficiente de oralidad conlleva una subrepresentación del constructo del conocimiento (Fernández Álvarez et al., 2022) que los estudiantes poseen y, muy directamente, ha favorecido la perpetuación de clases de Inglés "en español" (o en cualquier otra lengua autonómica).

Sin duda, todo lo expuesto anteriormente puede afectar a la interpretación y uso de las calificaciones de las pruebas y resta valor al estudiantado más preparado en inglés. El vacío legal y la falta casi total de un alfabetismo en evaluación supone que ninguno de los agentes

educativos reclame otra prueba más completa y exigente, aunque, obviamente, mucho más cara (Fernández Álvarez et al., 2022). Añadiremos que, además de la integración de las tareas orales para que aumente la representatividad del constructo de la prueba, se deben recopilar evidencias de validez empírica (Veas et al., 2020a), basadas en el uso real del idioma que requieren los estudiantes preuniversitarios para su incorporación a la Educación Superior.

En resumen, para garantizar la idoneidad de las tareas de la prueba se necesita más investigación empírica que nos ayude a comprender mejor el papel del desarrollo cognitivo y social/afectivo de los examinados en su desempeño en la prueba. Evidentemente una prueba que tiene tanta importancia debe recibir una mayor atención.

Tras lo anteriormente expuesto, el presente estudio tiene como objetivo general analizar la validez interna de la prueba de Lengua Extranjera-Inglés de la Evaluación para el Acceso a la Universidad diseñado y aplicado en la Comunidad de Madrid en el año 2021. Para ello, este trabajo plantea los siguientes objetivos específicos en relación a la prueba de Lengua Extranjera-Inglés:

1. Estudiar la calidad técnica.
2. Examinar las puntuaciones obtenidas por los sujetos en cada uno de los ítems.
3. Analizar las propiedades psicométricas de los ítems desde la Teoría Clásica de los Tests.
4. Estudiar la fiabilidad.
5. Examinar la distribución de los niveles de dificultad de los ítems.

3. Método

Se trata de un estudio con un diseño no experimental y transversal de carácter descriptivo en el que se utiliza la Teoría Clásica de Ítems y el modelo de Rasch.

3.1. Participantes

La población objeto de estudio está conformada por los aspirantes a la universidad que se presentaron a la prueba de Lengua Extranjera-Inglés, en convocatoria ordinaria, en la Comunidad de Madrid. Los datos, cedidos por una universidad pública de la Comunidad de Madrid, corresponden a 4,243 aspirantes a los estudios superiores universitarios en el año 2021. Los casos válidos para este estudio son 3,746, lo que representa en torno al 88.29% de los datos totales, garantizando la representatividad de la muestra. La pérdida muestral se debe a errores en la captación de las puntuaciones directas otorgadas a cada uno de los ítems.

3.2. Instrumento

El instrumento utilizado para este estudio es la prueba de lengua extranjera en la materia de Inglés diseñada en la Comunidad de Madrid en el año 2021[1]. Esta prueba está compuesta por un texto de comprensión lectora de una extensión aproximada de 300 palabras y por 13 ítems, los cuales pueden clasificarse en 5 bloques de tareas:

[1] La prueba usada en este estudio está disponible en la siguiente página web: https://www.ucm.es/file/inglÉs-3.

1. Parte 1: *True/False*. Se trata de dos ítems diseñados para medir la comprensión lectora. El estudiante debe decidir si dos frases que se le presentan son verdaderas o falsas, copiando a continuación únicamente el fragmento del texto que justifica su elección. Se otorga 1 punto por cada apartado y se califica con 0 puntos la opción elegida que no vaya justificada.
2. Parte 2: *Short answer*. Se trata de dos ítems que pretenden evaluar la comprensión lectora y la expresión escrita, mediante la formulación de dos preguntas abiertas que el estudiante debe contestar basándose en la información del texto, pero utilizando sus propias palabras en la respuesta. Cada una de las preguntas vale 1 punto, asignándose 0.5 puntos a la comprensión de la pregunta y del texto, y 0.5 a la corrección gramatical y ortográfica de la respuesta.
3. Parte 3: *Vocabulary*. Esta pregunta trata de medir el dominio del vocabulario en el aspecto de la comprensión. El estudiante debe demostrar esta capacidad localizando en el párrafo que se le indica un sinónimo, adecuado al contexto, de cuatro palabras o definiciones. Se adjudica 0.25 por cada apartado.
4. Parte 4: *Grammar*. Con esta pregunta se pretende comprobar los conocimientos gramaticales del estudiante, en sus aspectos morfológicos y/o sintácticos. Se presentan oraciones con huecos que el estudiante debe completar o rellenar. También pueden presentarse oraciones para ser transformadas u otro tipo de ítem. Se adjudica 0.25 a cada "hueco en blanco" y en el caso de las transformaciones o ítems de otro tipo se concede 0.5 con carácter unitario.
5. Parte 5: *Writing*. Se trata de una redacción, de 150 a 200 palabras, en la que el estudiante puede demostrar su capacidad para expresarse libremente en inglés. Se proponme una única opción y se otorgan 1.5 puntos por el buen dominio de la lengua (léxico, estructura sintáctica, etc.) y 1.5 por la madurez en la expresión de las ideas (organización, coherencia y creatividad). Para corregir esta redacción se utiliza la rúbrica de evaluación que se muestra en la Tabla 1.

Las preguntas de este examen fueron diseñadas por un comité formado por profesores de las distintas universidades de la Comunidad de Madrid especialistas en la materia de Inglés y se construyeron partiendo de los contenidos establecidos en el Real Decreto 1105/2014. 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.

3.3. Procedimiento

La prueba tuvo lugar el día 6 de junio de 2021 siguiendo las directrices establecidas para tal procedimiento. La prueba fue administrada en aulas habilitadas para tal fin. En cada grupo se establecieron las instrucciones de realización y cumplimentación de manera detallada informando a los aspirantes de que disponían de 90 minutos para contestar a la prueba. La muestra del estudio se examinó en una sola universidad de la Comunidad de Madrid y se utilizaron los resultados obtenidos de los distintos modelos de examen tanto en la prueba ordinaria como en la extraordinaria.

Las pruebas estadísticas se realizaron con el programa SPSS 28.0., así como con el *software* estadístico Winsteps v.4.4.0 (Linacre, 2019). En primer lugar, se llevaron a cabo análisis estadísticos descriptivos (medias, desviaciones típicas) de cada uno de los ítems. En segundo lugar, se analizó la fiabilidad de la prueba a través del índice de consistencia interna Alfa de Cronbach ordinal utilizando R con el paquete *psych* (Revelle, 2020). En tercer lugar, se realizaron análisis de frecuencias y porcentajes para determinar los índices de dificultad. Finalmente, se empleó el Modelo de Rasch para comparar las puntuaciones obtenidas por los examinandos en la prueba de Lengua Extranjera-Inglés, cuyas estimaciones se realizan mediante el método de máxima verosimilitud conjunta (Bond, 2004). Asimismo, se incorpora un Mapa de Wright que muestra la distribución de los niveles de dificultad de los ítems.

Este estudio se ha llevado a cabo bajo el cumplimiento de la normativa APA, de manera que se asegura la protección y el tratamiento de los datos de carácter personal y anonimato de los participantes quienes fueron identificados con un código.

4. Resultados

Los datos que aquí se presentan se corresponden con el total de estudiantes que se matricularon y, posteriormente, se presentaron a la prueba de Lengua Extranjera-Inglés en una universidad pública de la Comunidad de Madrid en el año 2021.

Por lo tanto, la Tabla 1 muestra los análisis descriptivos (medias y desviaciones típicas) de una muestra de 3,746 sujetos. Estos datos han permitido determinar la calidad técnica de la prueba de Lengua Extranjera-Inglés (objetivo específico 1).

Tabla 1. *Estadísticos descriptivos*

	Media	Desviación típica
True/False 1	.777	.416
True/False 2	.780	.415
Short answer 1	.690	.330
Short answer 2	.654	.358
Vocabulary 1	.186	.109
Vocabulary 2	.107	.124
Vocabulary 3	.135	.125
Vocabulary 4	.180	.113
Grammar 1	.301	.195
Grammar 2	.287	.194
Grammar 3	.298	.193
Grammar 4	.252	.250
Writing (Content)	.987	.453
Writing (Form)	.927	.466

Con relación a las puntuaciones obtenidas por los sujetos, así como las propiedades psicométricas de la prueba (objetivos específicos 2 y 3), se ha analizado la fiabilidad del instrumento a través del índice de consistencia interna Alfa de Cronbach ordinal. Se han obtenido valores

elevados (α =.864). Con la finalidad de determinar los índices de dificultad, se realizaron análisis de frecuencia y porcentaje de respuesta para cada uno de los ítems, según se muestra en la Tabla 2.

Tabla 2. *Frecuencia y porcentaje de respuesta para cada uno de los ítems*

Ítem	Puntuación	Frecuencia	Porcentaje (%)
True/False 1	0	836	22.3
	1	2910	77.7
True/False 2	0	826	22.1
	1	2920	77.9
Short answer 1	.00	399	10.7
	.25	201	5.4
	.50	829	22.1
	.75	792	21.1
	1.00	1525	40.7
Short answer 2	.00	531	14.2
	.25	307	8.2
	.50	724	19.3
	.75	692	18.5
	1.00	1492	39.8
Vocabulary 1	.00	966	25.8
	.25	2780	74.2
Vocabulary 2	.00	2136	57
	.25	1610	43
Vocabulary 3	.00	1719	45.9
	.25	2027	54.1
Vocabulary 4	.00	1057	28.2
	.25	2689	71.8
Grammar 1	.00	835	22.3
	.25	1313	35.1
	.50	1598	42.7
Grammar 2	.00	895	23.9
	.25	1405	37.5
	.50	1446	38.6
Grammar 3	.00	819	21.9
	.25	1386	37
	.50	1541	41.1
Grammar 4	.00	1855	49.5
	.50	1891	50.5
Writing (Content)	.00	198	5.3
	.25	241	6.4
	.50	421	11.2
	.75	530	14.1
	1.00	707	18.9
	1.25	607	16.2
	1.50	1042	27.8

	.00	258	6.9
	.25	299	8
	.50	500	13.3
Writing (Form)	.75	498	13.3
	1.00	684	18.3
	1.25	676	18
	1.50	831	22.2

Para estudiar la fiabilidad de la prueba (objetivo específico 4) se emplea el Modelo de Rasch, asumiéndose la posibilidad de comparar las puntuaciones obtenidas por los aspirantes en la prueba de Inglés. Se emplea el modelo de crédito parcial, cuyas estimaciones se realizan mediante el método de máxima verosimilitud conjunta. En este modelo, cada ítem es capaz de medir el constructo del rendimiento académico total en la prueba. En la Tabla 3 se muestran los índices de discriminación de la prueba.

Finalmente, con el objetivo 5 se plantea cómo es la distribución de los niveles de dificultad de los ítems que componen la prueba. Para llegar a conocerla, se calcularon los índices de ajuste de los ítems a través del Modelo de Rasch, según se muestra en la Tabla 3, que recoge los niveles de ajuste próximo (INFIT) y de ajuste lejano (OUTFIT) de cada ítem.

Tabla 3. *Nivel de dificultad e índices de ajuste de los ítems*

Ítem	PTBIS	Alfa de Cronbach si se elimina el ítem	Modelo de Rasch			
			Measure	ET	INFIT	OUTFIT
True/False 1	.391	.864	-.86	.04	1.09	1.20
True/False 2	.483	.859	-.87	.04	1.02	1.11
Short answer 1	.619	.853	-.09	.02	1.06	1.09
Short answer 2	.675	.850	.06	.02	1.06	1.02
Vocabulary 1	.595	.854	-.63	.04	.98	.90
Vocabulary 2	.536	.857	1.01	.04	.97	.96
Vocabulary 3	.435	.863	.45	.04	1.09	1.18
Vocabulary 4	.471	.861	-.48	.04	1.02	1.03
Grammar 1	.569	.855	.16	.03	1.11	1.12
Grammar 2	.602	.854	.30	.03	.97	.96
Grammar 3	.494	.860	.19	.03	1.08	1.10
Grammar 4	.517	.859	.63	.04	1.04	1.06
Writing (Content)	.767	.845	-.03	.01	.80	.77
Writing (Form)	.785	.844	.15	.01	.76	.74

Nota: PTBIS = Correlación punto biserial corregida; Measure = Parámetro de dificultad (*b*) del modelo de Rasch; ET = error típico asociado al parámetro de dificultad *b*; INFIT = Ajuste próximo; OUTFIT = Ajuste lejano.

El Mapa de Wright (Figura 3) muestra de manera más visual la distribución de los niveles de dificultad de los ítems. Este mapa, generado por el software estadístico Winsteps, muestra la distribución de los ítems a la derecha y a la izquierda la de personas. En la parte inferior aparecen los ítems más fáciles y arriba los más difíciles. Desde el modelo, se espera una distribución normal en las personas y una correspondencia entre los ítems y las personas de aproximadamente un 70%.

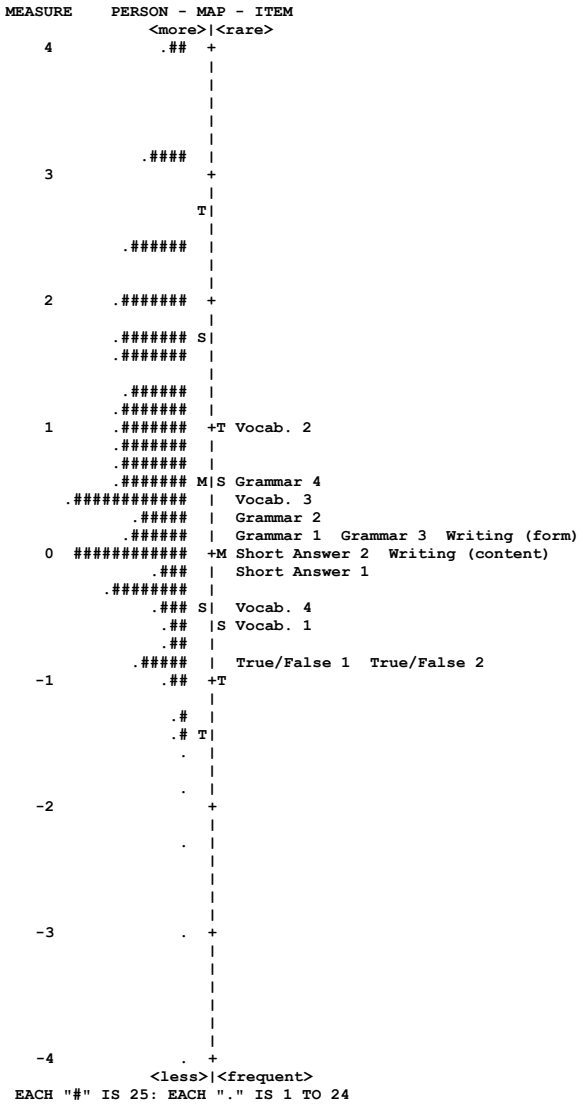

Figura 3. *Mapa de Wright de la prueba*

5. Discusión y conclusiones

Este estudio aporta evidencias sobre la calidad técnica de la prueba de Inglés de la Comunidad de Madrid que daba acceso a la universidad en España en el año 2021. Las conclusiones del estudio se exponen a continuación.

Con relación al objetivo 1, los estadísticos descriptivos muestran que los ítems de *True/False* podrían ser las preguntas que presentan mayores dificultades en término técnicos, ya que resultan ser los ítems más fáciles. Aunque los candidatos deben mostrar la evidencia de por qué la respuesta es verdadera o falsa copiando una frase del texto original, estos ítems no dejan de ser dicotómicos, y la respuesta es correcta o incorrecta. Por lo tanto, sería interesante revisar estos ítems, o incluso plantearse si deben incluirse en la prueba.

Por otro lado, cabe destacar el aspecto del contenido en el ejercicio de composición. Más de la cuarta parte de los candidatos (27.8 %) obtuvo la máxima puntuación (1.5) en este componente, tal y como se observa en la Tabla 2. Sin embargo, el porcentaje de candidatos que obtuvo esta máxima puntuación en aspectos formales es inferior (22.2 %). Esto nos hace plantearnos varias cuestiones. En primer lugar, habría que analizar si la rúbrica empleada está bien diseñada para tal efecto. Corregir un examen de alto impacto como el que se analiza en este estudio con una rúbrica deficiente pone en riesgo la fiabilidad y la validez de la prueba. Convendría mejorar dicha rúbrica e incluir descriptores más detallados que la transformen en una rúbrica más analítica, ya que el aspecto holístico no deja de darle cierta ambigüedad a la corrección, que acaba siendo bastante subjetiva. Por otro lado, habría que plantearse un estudio de la fiabilidad de los correctores que usan esta rúbrica. Estas son líneas para futuras investigaciones que pueden centrarse en (1) el tipo de formación recibida por los correctores para el uso de la rúbrica, (2) la consistencia interna de sus propias correcciones y (3) la fiabilidad entre correctores.

Respecto al objetivo 2, se observa que, en general, los candidatos tienen una buena puntuación en la mayoría de los ítems de la prueba, con excepción del segundo ejercicio de vocabulario, que resulta ser el ítem más difícil de toda la prueba, según se ve en la Tabla 2 y en la Figura 3. No es la finalidad de este estudio hacer un análisis cualitativo de los ítems, sino más bien abordar la calidad del examen desde una perspectiva más cuantitativa, y este dato nos pone en alerta sobre el comportamiento no solo de este ítem en particular sino también del ejercicio de vocabulario en general. El hecho de que haya dos ítems dentro del mismo componente con unos porcentajes elevados de candidatos con respuestas correctas (74.2% para el ejercicio 1 de vocabulario y 71.8% para el ejercicio 4) y otros dos ítems con unos porcentajes mucho más dispares (43% para el ejercicio 2 y 54.1% para el ejercicio 3), es indicativo de que el constructo de este ejercicio debería ser revisado. Es cierto que no todos los ítems deben tener el mismo índice de dificultad, pero cuando uno de los ítems de un mismo ejercicio se comporta de manera completamente opuesta al resto, este no sirve para discriminar entre los candidatos, ya bien sea porque el ítem es fácil o difícil. Esto nos hace cuestionarnos si la prueba pasó por un proceso de validación, y en el caso de que así fuese, en qué condiciones y qué revisiones se llegaron a hacer a los ítems que se comportaban de manera más problemática.

Cabe resaltar también, como se mencionó anteriormente, que en el ejercicio de composición los candidatos en general tienen una mayor puntuación en aspectos de contenido

que en cuestiones de forma, lo cual indica que o bien (1) los alumnos tienden a escribir de manera coherente y se adecúan al tema con opiniones originales, pero suelen tener errores gramaticales, de ortografía y de vocabulario, o que (2) los correctores suelen fijarse más en aspectos formales, al resultarle más fácil identificar ese tipo de errores. Según se indicó anteriormente, es una cuestión que merece una revisión más en profundidad.

En cuanto a los objetivos tres y cuatro, hay que resaltar que el nivel de fiabilidad de la prueba es bueno (Alfa de Cronbach ordinal =.864). Sin embargo, se recomendaría una revisión de los ítems de *True/False*, al tratarse de los dos ítems más fáciles de la prueba. Sin necesidad de analizar con más profundidad estos ítems, estos datos ya son indicio de que habría que hacer un estudio para (1) determinar si en realidad es necesario este componente en la prueba o, en caso de que se decida mantenerlo, (2) estudiar la posibilidad de modificarlo de tal manera que contribuya de manera más positiva a la fiabilidad de la prueba, ya que según se observa en la Tabla 3, el Alfa de Cronbach si se elimina el ítem 1, por ejemplo, no variaría. Sería necesario un replanteamiento de los ítems para que no sean la parte más fácil de la prueba, ya que el hecho de que los candidatos tengan que copiar una frase del texto como evidencia parece que no es la solución.

A pesar de las deficiencias expuestas anteriormente sobre el ejercicio de composición, es cierto que el nivel del Alfa de Cronbach bajaría a .845 y .844 respectivamente en caso de que se eliminaran las cuestiones de forma y contenido, lo cual indica que este ejercicio es el que mejor contribuye a la fiabilidad de la prueba. Los indicadores del constructo según los análisis realizados siguiendo el Modelo de Rasch demuestran que se pueden escalar en una sola dimensión y que todos los ítems registran un nivel de ajuste próximo (INFIT) y un nivel de ajuste lejano (OUTFIT) aceptable, según se muestra en la Tabla 3.

En relación al objetivo cinco, cabe mencionar que los ítems con mayor nivel de dificultad son el ejercicio 2 de *Vocabulary* (medida = 1.01 logits), el ejercicio 4 de *Grammar* (medida =.63 logits) y el ejercicio 3 de *Vocabulary* (medida =.45 logits). Por otro lado, los ítems más fáciles son el ejercicio 2 de *True/False* (medida =-.87 logits), el ejercicio 1 de *True/False* (medida =-.86 logits) y el ejercicio 1 de *Vocabulary* (medida =-.63 logits). Habría que subrayar también que según la información recogida en la Figura 2, la media de la mayor parte de los candidatos está por encima de la media los ítems, lo que puede ser indicio de que el nivel la prueba es inferior al nivel de los candidatos que se presentan a ella, que debe ser de B1 según el Marco Común de Referencia para las Lenguas (Council of Europe, 2001). Estos resultados demuestran que (1) nos tengamos que preocupar por la prueba en sí, que quizá deba someterse a un proceso de revisión en su totalidad para asegurarnos de que el constructo es el que debe ser, o que (2) el nivel de los alumnos que se presentan a este examen es superior al que tenían anteriormente como resultado de la implantación de programas bilingües en la Comunidad de Madrid. Sin embargo, no podemos llegar a esas conclusiones sin hacer un estudio del perfil de los candidatos que hacen la prueba. Desafortunadamente, ese es un dato que no se recoge en la prueba, y por tanto, una de las limitaciones de este estudio.

Todos los resultados aquí presentados van en la línea con los estudios que a día de hoy se han realizado sobre la prueba de Inglés para el acceso a la universidad (Fernández Álvarez, 2007; García Laborda et al., 2011; Amengual-Pizarro & Méndez García, 2012; García Laborda & Martín-Monje, 2013; Sevilla-Pavón, Gimeno-Sanz & García Laborda, 2017; Ruiz-Lázaro et al., 2021; Ruiz-Lázaro, 2022). Como prospectiva de investigación en

esta línea, tras los resultados encontrados alcanza sentido ahondar en el estudio de un diseño sistemático del examen que se base en los principios básicos del diseño de pruebas que incluya una fase de validación de ítems con la finalidad de determinar el comportamiento de ítems determinados y su posible inclusión en la prueba. Este diseño de la prueba debe ir acompañado de una adecuada formación a los correctores que se tendrán que enfrentar a evaluar ejercicios de expresión escrita a través de una rúbrica que idealmente siga unos principios más acordes a los de una prueba de alto impacto como este examen.

Tal y como se ha comentado anteriormente, la principal limitación se debe a la temporalidad debido a que estos datos se corresponden únicamente con la convocatoria ordinaria para el acceso a la universidad en el año 2021. Sin embargo, un atenuante es que la muestra estudiada es representativa de la población objeto de estudio. Podría complementarse con otros estudios sobre otras convocatorias del mismo o de diferentes cursos académicos. Esto permitiría comparar los datos y poder ver resultados más concluyentes. Por otro lado, en la línea de los estudios de Ruiz-Lázaro et al. (2021) y Ruiz-Lázaro (2022), este podría replicarse con el resto de pruebas de Inglés de las diferentes comunidades autónomas y aportar evidencias sobre la influencia del diseño y la calidad técnica de cada una de las pruebas en las puntuaciones promedio obtenidas por los aspirantes en las mismas. Esto, a su vez, permitiría observar aquellas que son más fáciles y más difíciles en función de su diseño.

Finalmente, cabe mencionar también la gran deficiencia de esta prueba al no incluir un componente de comprensión y expresión oral (García Laborda, 2013; Fernández Álvarez et al., 2022), produciéndose consecuencias directas e indirectas sobre todo en la enseñanza de Bachillerato. Hasta que el examen de inglés de las pruebas de acceso a la universidad no incluya un componente oral, este seguirá manteniendo un constructo que no es representativo y que no se basa en los elementos lingüísticos recogidos en el Marco Común de Referencia para las Lenguas, que en su última revisión llegó a desarrollar cuestiones relacionadas con la mediación lingüística o la interacción online, entre otras. Si nuestro examen de acceso a la universidad no incluye una parte de expresión oral, aún nos queda un largo camino que recorrer hasta que otros aspectos más novedosos puedan incluso plantearse. Todo esto puede llegar a impedir la mejora global de la competencia en L2 de nuestros futuros estudiantes universitarios. Esta información no justificada en evidencia se suele usar con motivos políticos más que académicos y no académicos que muestran lo contrario (Hughes & Madrid, 2020; Ródenas, 2018).

6. Referencias

Alderson, J. C., Clapham, C. & Wall, D. (1995). *Language test construction and evaluation.* Cambridge University Press.

Amengual Pizarro, M. (2004). *Análisis de la fiabilidad en las puntuaciones holísticas en ítems abiertos*. Tesis doctoral, Universidad Complutense de Madrid.

Amengual Pizarro, M. (2005). Posibles sesgos en los resultados del examen de selectividad. En H. Herrero Soler & J. García Laborda (Eds.). *Estudios y criterios para una selectividad de calidad en el examen de inglés* (pp. 121-148). Editorial de la Universidad Politécnica de Valencia.

Amengual Pizarro, M. (2006). Análisis de la prueba de inglés de selectividad de la Universitat de les Illes Balears. *Ibérica, 11,* 29-59.

Amengual-Pizarro, M. & Méndez García, M. C. (2012). Implementing the oral English task in the Spanish university admission examination: An international perspective of the language. *Revista de Educación, 357*, 105-127.

Bachman, L. F. (1990). *Fundamental considerations in language testing.* Oxford University Press.

Bachman, L. F. & Palmer, A. S. (1996). *Language testing in practice: Designing and developing useful language tests.* Oxford University Press.

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

Brown, H. D. (1994). *Principles of language learning and teaching*. Prentice Hall.

Bueno, M. C. & Luque, G. (2012). Competencias en lengua extranjera exigibles en la Prueba de Acceso a la Universidad: Una propuesta para la evaluación de los aspectos orales. *Revista de Educación, 357*, 81-104.

Carr-Hill, R., Rolleston, C., Schendel, R. & Waddington, H. (2018). The effectiveness of school-based decision making in improving educational outcomes: A systematic review. *Journal of Development Effectiveness, 10*(1), 61-94. https://doi.org/10.1080/19439342.2018.1440250.

Council of Europe. (2001). Common European framework of reference for language learning, teaching and assessment. Cambridge University Press.

CRUE (2022). *Propuesta de prueba de acceso a la Universidad a partir del curso 2023-2024*. Ministerio de Educación y Formación Profesional. https://www.educacionyfp.gob.es/en/dam/jcr:fe46dfb9-07a5-4482-8ebb-e070f157de77/presentaci-n-simplificada-evau.pdf.

Díez Bedmar, M. B. (2011). The English exam in the university entrance examination: An overview of studies. *Revista Canaria de Estudios Ingleses, 63*, 101-112.

Fernández Álvarez, M. (2007). Propuesta metodológica para la creación de un nuevo examen de inglés en las pruebas de acceso a la universidad. Tesis doctoral, Universidad de Granada.

Fernández Álvarez, M., García Laborda, J. & Magal-Royo, T. (2022). Underrepresentation of the construct in Spanish standardized foreign language exams: A computer assisted exam proposal, *Porta Linguarum,* Monograph IV, 27-45. https://doi.org/10.30827/portalin.vi.21393.

Fulcher, G. (1997). An English language placement test: Issues in reliability and validity. *Language Testing, 14*(2), 113-139.

García Laborda, J. (2010). ¿Necesitan las universidades españolas una prueba de acceso informatizada? El caso de la definición del constructo y la previsión del efecto en la enseñanza para idiomas extranjeros. *Revista de Orientación y Psicopedagogía, 21*(1), 71-80.

García Laborda, J. (2013). Reacciones iniciales de los profesores a la preparación de la prueba informatizada de acceso a la universidad. *Lenguaje y Textos, 38*, 133-140.

García Laborda, J. & Gimeno Sanz, A. (2008). Adaptación del examen de inglés de las pruebas de acceso a la universidad a un entorno informático: Estudio sobre la tipología de preguntas. *Proceedings of the XXV Congreso Nacional de Lingüística* Aplicada (pp. 723-730). Servicio de publicaciones de la Universidad de Murcia.

García Laborda, J. & Martín-Monje, E. (2013). Item and test construct definition for the new Spanish Baccalaureate final evaluation: A proposal. *International Journal of English Studies, 13*(2), 69-88. https://doi.org/10.6018/ijes.13.2.185921.

García Laborda, J., Gimeno Sanz, A. & de Siqueira, J. M. (2011). Experimentación de las soluciones tecnológicas del proyecto PAULEX para optimizar la prueba de inglés del examen de acceso a la universidad en España. *Educaçao Temática Digital, 12*(2), 1-11.

Hughes, A. (2003). *Testing for language teachers*. Cambridge University Press.

Hughes, S. P. & Madrid, D. (2020) The effects of CLIL on content knowledge in monolingual contexts. *The Language Learning Journal, 48*(1), 48-59. https://doi.org/10.1080/09571736.2019.1671483.

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

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, 3*, de 3 enero de 2015, pp. 169-546.

Real Decreto-ley 5/2016, de 9 de diciembre, de medidas urgentes para la ampliación del calendario de implantación de la Ley Orgánica 8/2013, de 9 de diciembre, para la mejora de la calidad educativa. *Boletín Oficial del Estado, 298*, de 10 de noviembre de 2016, pp. 86168-86174.

Real Decreto 310/2016, de 29 de julio, por el que se regulan las evaluaciones finales de Educación Secundaria Obligatoria y de Bachillerato. *Boletín Oficial del Estado, 183*, de 30 de julio de 2016, pp. 53049-53065.

Real Decreto 243/2022, de 5 de abril, por el que se establecen la ordenación y las enseñanzas mínimas del Bachillerato. *Boletín Oficial del Estado, 82,* de 6 de abril de 2022, pp. 1-325.

Revelle, W. (2020). *psych: Procedures for Personality and Psychological Research*. Northwestern University.

Ródenas, J. A. (2018). El impacto de la enseñanza bilingüe en el alumnado de Educación Primaria: análisis del rendimiento académico y de los intereses por las áreas curriculares. *Ensayos, Revista de la Facultad de Educación de Albacete, 33*(2), 1-13. https://doi.org/10.18239/ensayos.v33i2.1532.

Ruiz-Lázaro, J. (2022). *Acceso a la universidad en España. Análisis comparativo de las pruebas comunes por comunidades autónomas* [Tesis doctoral, Universidad Complutense de Madrid] https://eprints.ucm.es/id/eprint/70697/.

Ruiz-Lázaro, J., & González Barbera, C. (2017). Análisis de la Prueba de Lengua Castellana y Literatura que da acceso a la universidad. Comparación entre las comunidades autónomas. *Bordón, 69*(3), 175-195. https://doi.org/10.13042/Bordon.2017. 50927.

Ruiz-Lázaro, J., González Barbera, C., & Gaviria Soto, J.L. (2021). Las pruebas de inglés para acceder a la Universidad. Una comparación entre Comunidades Autónomas. *Educación XX1, 24*(1), 233-270. http://doi.org/10.5944/educXX1.26746.

Sevilla-Pavón, A., Gimeno-Sanz, A., & García Laborda, J. (2017). Actitudes docentes hacia los ejercicios de la Prueba de Acceso a la Universidad informatizada. *Educaçao e Pesquisa: Revista da Faculdade de Educação da Universidade de São Paulo, 43*(4), 1179-1200. https://doi.org/10.1590/S1517-9702201612149283.

Sistema Integrado de Información Universitaria (2022). *Estadísticas de las Pruebas de Acceso a la Universidad*. Ministerio de Educación, Cultura y Deporte.

Veas, A., Benítez, I., Navas, L., & Gilar-Corbí, R. (2020a). Comparative analysis of the University Entrance Examinations within the construct comparability approach. *Revista de Educación, 388,* 65-84. https://doi.org/10.4438/1988-592X-RE-2020-388-447.

Veas, A., Navas, L., Pozo-Rico, T., & Miñano, P. (2020b). University Entrance Examinations in Spain: Using the Construct Comparability Approach to Analyze Standards Quality. *Frontiers in Psychology, 11*: 127. https://doi.org/10.3389/fpsyg.2020.00127.