



Efectos del formato del ítem, la familia lingüística y tipología del sistema ortográfico en la comprensión lectora en L2

Effects of Item Format, Language Family, and Orthographic System Typology on L2 Reading Comprehension

Karen Martínez

Algaciras 

UNIVERSIDAD DEL PAÍS VASCO
ESPAÑA

kmartinez019@ikasle.ehu.eus

Paula Elosua 

UNIVERSIDAD DEL PAÍS VASCO
ESPAÑA

paula.elosua@ehu.eus

Javier Fruns

Giménez 

INSTITUTO CERVANTES
ESPAÑA

javier.fruns@cervantes.es

Recibido: 11-4-2025 / **Aceptado:** 6-7-2026

DOI: 10.4151/S0718-09342026012101495

Resumen

Garantizar la comparabilidad de las puntuaciones constituye un aspecto central en las evaluaciones de comprensión lectora en L2, especialmente cuando diferentes fuentes de variabilidad pueden afectar la interpretación de los resultados. En este contexto, el presente estudio examina la relación entre el formato del ítem y determinadas propiedades lingüísticas de la L1 —como la transparencia ortográfica, la familia tipológica y el sistema de escritura—, con el objetivo de analizar cómo varía el rendimiento en función de estas características lingüísticas. Para ello, se analizaron los datos de 8.701 participantes en el examen de certificación DELE B2, mediante modelos lineales mixtos con medidas repetidas. Los hablantes de lenguas transparentes, alfabéticas e indoeuropeas tendieron a obtener puntuaciones más altas, aunque con variaciones según el tipo de tarea. Los formatos de opción múltiple y cloze selectivo acentuaron algunas diferencias entre grupos, mientras que el cloze racional presentó patrones de desempeño más homogéneos. En conjunto, los resultados sugieren que la relación entre el formato del ítem y el rendimiento varía en función de las características lingüísticas de la L1. Se discuten las implicaciones de estos resultados sobre la validez de las puntuaciones en contextos multilingües, destacando la necesidad de considerar el formato como un factor de método.

Palabras clave: comprensión lectora en L2, formato del ítem, distancia lingüística, sistema de escritura, transparencia ortográfica

Abstract

Ensuring score comparability constitutes a central issue in L2 reading comprehension assessment, particularly when multiple sources of variability may affect the interpretation of test results. Within this context, the present study examines the relationship between item format and specific L1 linguistic properties—such as orthographic transparency, typological family, and writing system—in order to analyze how performance varies as a function of these linguistic characteristics. To this end, data from 8,701 participants who took the DELE B2 certification examination were analyzed using repeated-measures linear mixed models. Speakers of transparent, alphabetic, and Indo-European languages tended to obtain higher scores, although the magnitude of these differences varied across task types. Multiple-choice and selective cloze formats accentuated some between-group differences, whereas rational cloze showed more homogeneous performance patterns. Overall, the findings suggest that the relationship between item format and performance varies according to L1 linguistic characteristics. The implications of these findings for score validity in multilingual contexts are discussed, highlighting the need to consider item format as a potential method factor.

Keywords: L2 reading comprehension, item format, language distance, writing system, orthographic transparency

INTRODUCCIÓN

La comprensión lectora en una segunda lengua (L2) es una competencia compleja y multidimensional en la que interactúan de manera dinámica factores lingüísticos y cognitivos (Grabe, 2009). Las propiedades de la lengua materna (L1), por ejemplo, se han asociado con diferencias en el desempeño lector y en las demandas implicadas en el reconocimiento y la integración de la información escrita (Frost, 2012; Landerl et al., 2022; Verhoeven & Perfetti, 2022). La literatura especializada ha destacado tres dimensiones lingüísticas relevantes para la lectura en L2: a) la transparencia ortográfica, que influye la eficiencia de la decodificación en función del grado de consistencia en la correspondencia grafema-fonema (Katz & Frost, 1992; Pae, 2018; Shepperd, 2024); b) el sistema de escritura, que modula las demandas de procesamiento visual y lingüístico implicadas en la lectura según su complejidad gráfica (Hemelstrand & Inoue, 2024); y c) la distancia lingüística entre la L1 y la L2, cuya proximidad tipológica y funcional facilita la transferencia de conocimientos y estrategias lectoras entre ambas lenguas (Jeon & Yamashita, 2014; Schepens et al., 2016; Chai & Bao, 2023).

La complejidad inherente a la comprensión lectora en L2 como constructo, plantea asimismo desafíos para su evaluación. La investigación ha mostrado que el formato del ítem constituye un componente central del proceso de medida, en tanto condiciona el tipo de respuesta requerida y, en consecuencia, las inferencias derivadas sobre el desempeño (Messick, 1989; Bachman & Palmer, 1996; Becker & Nekrasova-Becker, 2018; Alderson, 2000). Este aspecto adquiere especial relevancia en evaluaciones multilingües, donde el tipo de tarea puede interactuar con las

características de la lengua materna (Kobayashi, 2002b; Liao, 2023). En este sentido, la literatura ha señalado que distintos formatos de ítem pueden asociarse con diferentes formas de abordar las tareas de lectura —como la lectura cuidadosa, el escaneo o el uso de atajos—, lo cual podría influir en el desempeño observado (Brunfaut & McCray, 2015).

En este marco, el presente estudio investigó el efecto conjunto de tres propiedades lingüísticas de la L1 (transparencia ortográfica, familia lingüística y sistema de escritura) y el formato del ítem sobre el desempeño en una prueba de comprensión lectora en español como L2. Se analizaron las respuestas de 8.701 participantes en un examen oficial de acreditación de nivel B2. Los resultados mostraron que los hablantes de lenguas con ortografía transparente, sistemas de escritura alfabéticos y pertenecientes a familias indoeuropeas tendieron a obtener mejores resultados, aunque estas ventajas no se observaron de forma uniforme en todas las tareas. En particular, los ítems cloze selectivo tendieron a presentar una mayor dificultad relativa. En conjunto, el trabajo concluye que la relación entre el formato del ítem y el rendimiento puede variar en función del perfil lingüístico de los evaluados, lo cual plantea implicaciones para la comparabilidad de las puntuaciones en evaluaciones de comprensión lectora en L2.

1. Marco teórico

1.1 Comprensión lectora en L2

La comprensión lectora (CL) es un proceso complejo y multifacético que integra componentes interrelacionados (Grabe, 2009). Son varias las propuestas teóricas que intentan explicar la forma en que los lectores procesan la información y los factores que inciden en su desempeño (p. ej., Gernsbacher, 1990; Graesser et al., 1994; Kintsch, 1988; Myers & O'Brien, 1998; van den Broek et al., 1999; Perfetti & Stafura, 2014; Verhoeven & Perfetti, 2022). La CL suele describirse como el resultado de la interacción entre habilidades de bajo nivel —como la decodificación y el acceso a representaciones fonológicas, ortográficas, semánticas y sintácticas— y habilidades de alto nivel —como la integración de ideas, el reconocimiento de la estructura textual y de las ideas principales, la identificación de la intención del autor y la generación de inferencias para construir una representación coherente del texto— (Kintsch, 1988; Balota et al., 1990; Perfetti & Stafura, 2014; Graesser, 2015). Estas habilidades no se desarrollan de manera idéntica en todas las lenguas pues están condicionadas en parte, por las características del sistema de escritura y las propiedades ortográficas de cada una de ellas. Aspectos como la correspondencia grafema-fonema, la complejidad ortográfica y las demandas perceptivas del sistema escrito pueden influir en la velocidad y eficacia con que los lectores adquieren y automatizan los procesos implicados en la comprensión (Frost, 2012; Pae, 2018; Landerl et al., 2022; Verhoeven & Perfetti, 2022).

1.2 El impacto del sistema de escritura y ortografía en la comprensión lectora en L2

El sistema de escritura y la ortografía inciden en el rendimiento lector en L2, en la medida en que modulan la decodificación, la fluidez y el acceso al léxico (McBride et al., 2022; Verhoeven & Perfetti, 2022; Shepperd, 2024). Desde una perspectiva tipológica, los sistemas de escritura pueden clasificarse según su correspondencia con la fonología; los alfabéticos (p. ej., español) representan fonemas; los silábicos (p. ej., hiragana japonés) codifican sílabas; y los logográficos o morfossilábicos (p. ej., chino o kanji japonés) asocian caracteres con unidades léxicas o morfélicas. Estos últimos exigen mayores demandas de procesamiento (Sproat & Gutkin, 2021; Hemelstrand & Inoue, 2024). El japonés ilustra esta complejidad al combinar tres sistemas con funciones complementarias: *hiragana* y *katakana*, de carácter transparente, y *kanji*, altamente opaco, que implica memoria visual y conocimiento morfológico (Inoue et al., 2017; Wang et al., 2024).

En ortografías transparentes, como el español o el finlandés, la correspondencia sistemática entre grafemas y fonemas favorece una automatización más rápida y consistente de la lectura. En contraste, en ortografías opacas, como el inglés o el francés, las irregularidades ortográficas incrementan las demandas del procesamiento lector y favorecen una mayor dependencia de estrategias compensatorias, como el reconocimiento léxico global o el análisis morfológico (Pae, 2018; Shepperd, 2024; Zhang et al., 2025). En este contexto, las diferencias tipológicas entre sistemas de escritura y ortografías pueden constituir un elemento relevante para interpretar posibles variaciones en el rendimiento lector en L2.

1.3 Influencia de la distancia lingüística en la comprensión lectora en L2

La distancia lingüística, entendida como el grado de diferencia tipológica entre L1 y L2 en aspectos como gramática, morfología, fonología, sintaxis, semántica y sistema de escritura (Radman et al., 2021; Cargnelutti et al., 2022), incide en el desarrollo de la competencia lectora en L2. Desde el modelo de la competencia subyacente común y la hipótesis de la interdependencia (Cummins, 1979, 1981), se plantea que ambas lenguas comparten una base cognitiva, de modo que ciertas habilidades desarrolladas en la L1 pueden transferirse a la L2. En este sentido, se ha concluido que una mayor similitud tipológica favorece la transferencia facilitando la comprensión lectora en L2 (Jeon & Yamashita, 2014; Schepens et al., 2016; Zhang & Zhang, 2022).

No obstante, la transferencia interlingüística no depende exclusivamente de la cercanía tipológica, sino también del grado de correspondencia estructural y funcional entre sus sistemas lingüísticos (Muñoz et al., 2018; Chai & Bao, 2023). Cuando estructuras gramaticales, sintácticas o léxicas cumplen funciones similares en ambas lenguas, la transferencia tiende a verse favorecida; en cambio, cuando dichas

correspondencias son limitadas, el apoyo potencial de la L1 durante la comprensión en L2 se reduce.

Desde una perspectiva cognitiva, la investigación sobre activación interlingüística ha mostrado que la L1 puede permanecer activa incluso durante la lectura exclusiva en L2 (Bobb et al., 2020). Cuando ambas lenguas comparten similitudes tipológicas, esta coactivación puede favorecer el reconocimiento léxico y el acceso al significado. Por el contrario, una mayor distancia lingüística puede incrementar las dificultades para integrar información léxica y sintáctica durante la comprensión en L2 (Hao et al., 2024; Yamashita, 2008).

1.4 Evaluación de la comprensión lectora en contextos multilingües: validez, formato del ítem y comparabilidad de las puntuaciones

En la evaluación de la comprensión lectora en L2, el análisis del formato del ítem debe situarse dentro de marcos contemporáneos de validación. Desde la visión unificada de la validez (Messick, 1989, 1995) y el enfoque argumentativo de Kane (2013), las puntuaciones deben respaldar interpretaciones e inferencias coherentes con el constructo evaluado, considerando tanto las características de las tareas como las condiciones bajo las cuales se obtienen las respuestas. En esta línea, los *Standards for Educational and Psychological Testing* (AERA, APA, & NCME, 2014) subrayan la necesidad de justificar explícitamente cómo se aplican, interpretan y utilizan los resultados de una prueba.

Los enfoques sociocognitivos de la evaluación (Bachman & Palmer, 1996; Weir, 2005) amplían esta perspectiva al considerar que los ítems no son simples contenedores de contenido, sino unidades de evaluación cuyo diseño puede condicionar la manera en que los examinados interactúan con la tarea (Brunfaut & McCray, 2015; McCray & Brunfaut, 2018). Desde este enfoque, el formato del ítem constituye un componente relevante de la representación operativa del constructo.

Estas consideraciones adquieren especial relevancia en contextos multilingües, donde el formato del ítem puede interactuar con variables lingüísticas como la L1. En estos casos, las diferencias en el desempeño podrían reflejar no solo variaciones en comprensión lectora, sino también distintos grados de ajuste entre las demandas de la tarea y el perfil lingüístico de los evaluados (Liao, 2023). Desde el marco de la validez, identificar estos efectos resulta relevante para fundamentar decisiones de diseño y para la interpretación y comparabilidad de las puntuaciones entre grupos (Messick, 1989; Ureña et al., 2024; Weir, 2005).

En consonancia con estos planteamientos, diversas instituciones dedicadas a la certificación lingüística como Cambridge University Press & Assessment (2022), el Goethe-Institut (2018), la Università per Stranieri di Perugia (2017) o el Instituto

Cervantes (2012), incorporan formatos de tarea diferentes en sus pruebas de comprensión lectora incluyendo ítems de opción múltiple, cloze o emparejamiento, con el fin de fortalecer las evidencias de validez basadas en procesos (Bachman & Palmer, 1996; O'Sullivan & Weir, 2011).

La investigación ha mostrado que los distintos formatos de ítem se asocian con estrategias de respuesta y perfiles de procesamiento diferenciados. Los ítems de opción múltiple, caracterizados por altos niveles de fiabilidad y discriminación (Haladyna & Rodríguez, 2013; Becker & Nekrasova-Beker, 2018), se han vinculado con estrategias de reconocimiento, localización o escaneo superficial (Ozuru et al., 2013; Lim, 2014, 2019). En contraste, los ítems tipo cloze —incluidas variantes como el cloze racional y el selectivo— suelen requerir una mayor integración léxica, sintáctica y semántica, movilizando procesos de coherencia e inferencia (Jonz, 1976; Oller & Jonz, 1994; Kobayashi, 2002a, 2002b; Brown, 2013; McCray & Brunfaut, 2018). Por su parte, los ítems de emparejamiento evalúan la integración de información a nivel global, aunque su desempeño depende en gran medida de la calidad del diseño y del control de las pistas superficiales (Zeigenfuse et al., 2020).

1.5 Preguntas de investigación

A partir de los antecedentes revisados, este estudio examina la relación entre el formato del ítem, las propiedades lingüísticas de la L1 y el desempeño en comprensión lectora en español como L2. En particular, se pretende analizar si determinadas características lingüísticas de los examinados se asocian con patrones diferenciales de rendimiento según el tipo de tarea utilizada en la evaluación:

- PI1. ¿Se observan diferencias en el desempeño en comprensión lectora en español como L2 en función del formato de los ítems utilizados en la prueba?
- PI2. ¿Se asocia el desempeño en comprensión lectora en español como L2 con las propiedades lingüísticas de la lengua materna (familia tipológica, transparencia ortográfica y sistema de escritura)?
- PI3. ¿Existen efectos de interacción entre el formato del ítem y las propiedades lingüísticas de la L1 en el desempeño en comprensión lectora en español como L2?

2. Marco metodológico

2.1 Participantes

La muestra estuvo compuesta por 8.701 candidatos que realizaron una prueba de comprensión lectora de español como segunda lengua para la acreditación del nivel B2 del Marco Común Europeo de Referencia para las Lenguas (MCER; Consejo de Europa, 2002). La aplicación de la prueba tuvo lugar en 81 países. Los datos,

suministrados de manera anonimizada por el Instituto Cervantes, corresponden a una base secundaria sin información sobre país de nacimiento ni lengua materna declarada. Por esta razón, se empleó el país de aplicación como proxy de la lengua predominante de los participantes. A partir de esta asignación contextual, las lenguas se clasificaron en función de tres dimensiones analíticas: (a) grado de transparencia u opacidad ortográfica; (b) sistema de escritura (Sproat & Gutkin, 2021); y (c) familia lingüística. Esta operacionalización permite examinar patrones generales de desempeño considerando los contextos tipológicos asociados a las lenguas predominantes en los países de aplicación (véase Tabla 1).

Tabla 1

Tipología lingüística y puntuaciones por formato de ítem

Variable de agrupación	Categoría	N	%	T1 Opción múltiple	T2 Correspon- dencia	T3 Cloze racional	T4 Cloze selectivo	Puntuación final
Transparencia ortográfica	Transparente	7.452	85.65	4.06 (1.33)	7.43 (1.65)	4.13 (1.63)	11.05 (1.98)	26.69 (4.57)
	Opaca	1.249	14.35	3.60 (1.41)	6.91 (1.99)	3.47 (1.82)	9.49 (2.34)	23.49 (5.55)
Total		8.701	100					
Sistema de escritura	Alfabético	8.330	95.74	4.03 (1.34)	7.41 (1.67)	4.09 (1.65)	10.92 (2.04)	26.46 (4.67)
	Logográfico/silábico	136	1.56	2.97 (1.29)	5.42 (2.0)	2.53 (1.64)	7.91 (2.16)	18.84 (4.96)
	Abjad	173	1.99	3.27 (1.37)	6.58 (2.20)	2.97 (2.01)	9.43 (2.55)	22.27 (5.79)
	Logográfico	60	0.69	3.00 (1.34)	6.30 (2.06)	3.35 (1.81)	9.00 (2.36)	21.65 (5.89)
	Abugida	2	0.02	5.00 (0.00)	8.50 (0.70)	4.50 (2.12)	13.00 (0.00)	31.00 (2.83)
Total		8.701	100					
Familia lingüística	lenguas indoeuropeas	8.133	93.47	4.05 (1.33)	7.43 (1.65)	4.11 (1.64)	10.97 (2.01)	26.57 (4.60)
	lenguas japónés-coreanas	212	2.44	2.98 (1.33)	5.69 (2.08)	2.87 (1.67)	8.24 (2.15)	19.78 (5.12)
	lenguas afroasiáticas-semíticas	177	2.03	3.26 (1.36)	6.59 (2.18)	2.93 (2.03)	9.41 (2.55)	22.21 (5.77)
	lenguas sino-tibetanas	63	0.72	3.06 (1.34)	6.33 (2.04)	3.39 (1.80)	9.12 (2.39)	21.92 (5.88)
	lenguas jemer-tailandesas	42	0.48	3.47 (1.01)	6.66 (1.64)	3.16 (1.77)	9.45 (1.86)	22.76 (4.66)
	lenguas turcas-altaicas	29	0.33	3.00 (1.43)	6.27 (2.41)	3.24 (1.97)	8.34 (2.19)	20.86 (6.18)
	lenguas niger-congo	45	0.52	3.60 (1.45)	7.00 (1.91)	2.93 (1.69)	8.93 (3.00)	22.46 (4.95)
Total		8.701	100	4.00 (1.35)	7.35 (1.71)	4.04 (1.67)	10.83 (2.10)	26.23 (4.85)

Nota. Entre paréntesis se indican las desviaciones estándar. Los resultados de la categoría Abugida (N = 2) no se incluyeron en los análisis por sistema de escritura debido a la baja representatividad de la muestra. Sin embargo, sí fueron considerados en los análisis por familia lingüística, integrándose en el grupo de lenguas jemer-tailandesas (N = 42).

2.1.1. Consideraciones éticas

El estudio utilizó exclusivamente una base de datos secundaria completamente anonimizada y sin información personal identificable. Su uso se ajusta a las directrices internacionales para la utilización responsable de datos en investigación educativa y a las normativas institucionales aplicables. Debido al tipo de dato, el estudio está exento de requerir consentimiento informado adicional o aprobación por un comité de ética.

2.2 Diseño de investigación

El estudio adopta un diseño ex post facto, de tipo comparativo y correlacional, en el que no hubo manipulación de variables independientes. La variable dependiente fue

la puntuación obtenida en la prueba de comprensión lectora. Como variables independientes, se incluyeron: (a) ortografía (transparente y opaca); (b) familia lingüística de la L1 (indoeuropeas, afroasiáticas-semíticas, japonés-coreanas, sino-tibetanas, jemer-tailandesas, níger-congo y turcas-altaicas); (c) sistema de escritura (alfabético, logográfico/silábico, abjad, logográfico); y (d) formato del ítem (opción múltiple, correspondencias, cloze racional y cloze selectivo).

2.3 Materiales y procedimiento

Prueba de comprensión lectora del examen DELE (Diploma de Español como Lengua Extranjera—Instituto Cervantes), nivel B2. Consta de cuatro tareas con un total de 36 ítems: a) Tarea 1 (T1-Ítems de opción múltiple, 6 ítems), b) Tarea 2 (T2-Ítems de correspondencias, 10 ítems), c) Tarea 3 (T3-Ítems cloze racional, 6 ítems); d) Tarea 4 (T4 -Ítems cloze selectivo, 14 ítems). El rango de la puntuación final es [0, 36]. Un ejemplo de prueba puede consultarse en <https://exámenes.cervantes.es/es/dele/exámenes/b2>.

DELE B2 es un diploma oficial administrado por el Instituto Cervantes, reconocido internacionalmente y enmarcado en el Marco Común Europeo de Referencia para las Lenguas (MCER). El Instituto Cervantes, como miembro de Association of Language Testers in Europe (ALTE), garantiza que la administración y calificación de estas pruebas se realicen bajo estándares de calidad aceptados por la comunidad internacional, asegurando consistencia, rigor científico y validez en sus procesos (ALTE, 2024).

2.4 Análisis de datos

El análisis se desarrolló en dos fases. En la primera, se evaluó la calidad psicométrica de la prueba mediante un análisis factorial confirmatorio (AFC) que comparó la idoneidad de un modelo unidimensional y un modelo jerárquico de segundo orden. La evaluación del ajuste siguió los criterios de Hu y Bentler (1999). Se estimó además la consistencia interna mediante el coeficiente omega de McDonald (2013).

En la segunda fase, se examinaron las diferencias de rendimiento en función del formato del ítem y de las variables lingüísticas mediante modelos lineales mixtos (LMM), dada su idoneidad para datos con medidas repetidas, tamaños muestrales desbalanceados y posibles heterocedasticidades entre grupos (Baayen, et al., 2008; Luke, 2017; West et al., 2014). Los modelos incluyeron como efectos fijos el formato del ítem, las variables lingüísticas y sus interacciones, mientras que la variabilidad asociada a los participantes se modeló como efecto aleatorio, con el fin de considerar la dependencia intra-sujeto. Para cada modelo se reportaron las pruebas F correspondientes a los efectos fijos y los tamaños del efecto mediante eta cuadrado parcial (η^2p) calculado a partir de tablas ANOVA Tipo III con aproximación de

Kenward–Roger. Asimismo, se estimaron medias marginales (*estimated marginal means*) y, en los análisis con múltiples grupos, se realizaron comparaciones post hoc con ajuste de Tukey. Cuando fue pertinente, las diferencias entre grupos se representaron mediante *compact letter display* (CLD; Piepho, 2004), donde los grupos que comparten al menos una letra no difieren significativamente entre sí. Con fines interpretativos, se calcularon tamaños del efecto (Cohen's *d*. 1988) para cada par de comparaciones.

Con el fin de evaluar la estabilidad de los resultados frente al desbalance muestral entre grupos, se llevó a cabo un procedimiento de validación cruzada, a partir de la muestra original se extrajo una muestra aleatoria equilibrada en función del tamaño del grupo en la que se replicaron todos los análisis.

3. Resultados

3.1 Propiedades psicométricas

El modelo unidimensional presentó un ajuste deficiente ($\chi^2 = 6018.94$, g.l. = 594, $p < .001$; CFI = .895; TLI = .889; RMSEA = .032; SRMR = .056), mientras que el modelo jerárquico de segundo orden con un factor general de comprensión lectora mostró un ajuste adecuado ($\chi^2 = 4264.13$, g.l. = 590, $p < .001$; CFI = .938; TLI = .934; RMSEA = .027; SRMR = .043) con cargas elevadas de las tareas sobre el factor general (T1 = .82, T2 = .91, T3 = .70, T4 = .73). La consistencia interna del factor general estimada por el coeficiente omega fue .88.

3.2 Comparaciones por transparencia ortográfica

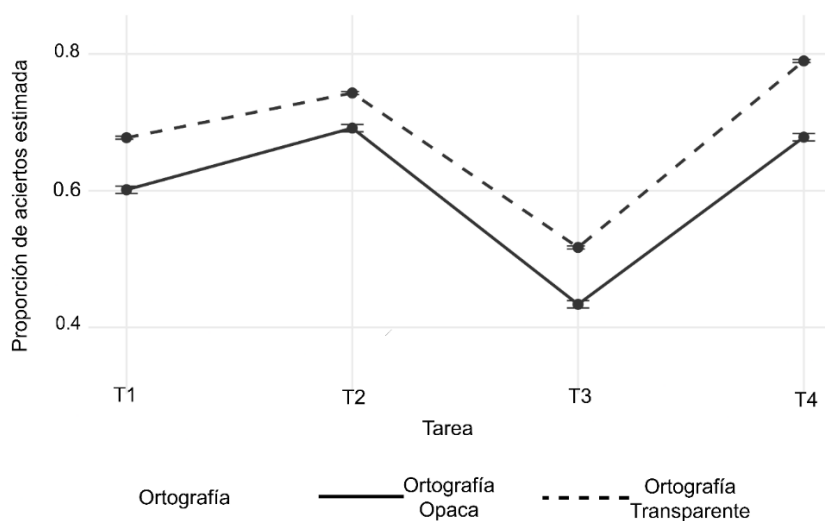
Se identificaron efectos principales significativos para la ortografía, $F(1, 8699) = 405.13$, $p < .001$, $\eta^2p = .04$, y para el tipo de tarea, $F(3, 26097) = 2382.77$, $p < .001$, $\eta^2p = .22$. La interacción entre ambos factores también fue significativa, $F(3, 26097) = 26.15$, $p < .001$, aunque con un tamaño del efecto muy reducido ($\eta^2p = .003$).

El análisis de las medias marginales estimadas concluyó que, en las cuatro tareas, los hablantes de lenguas con ortografía transparente obtuvieron mejores resultados promedio que aquellos con ortografías opacas (Gráfico 1). Traduciendo las diferencias a número de ítems respondidos correctamente, la ventaja del grupo transparente es de aproximadamente 0.46 ítems en T1 (6 ítems), 0.51 ítems en T2 (10 ítems), 0.66 ítems en T3 (8 ítems) y 1.56 ítems en T4 (14 ítems), siendo todas las diferencias estadísticamente significativas ($p < .001$). Los tamaños del efecto (Cohen's *d*) fueron moderados en T1 ($d = 0.48$) y en T3 ($d = 0.53$), algo menores T2 ($d = 0.32$) y grandes en T4 ($d = 0.70$). En cuanto al patrón de dificultad de las tareas, para ambos grupos T3 fue la tarea más difícil, mientras que T2 y T4 resultaron más sencillas. Aunque la interacción entre transparencia ortográfica y tipo de tarea presentó un tamaño del efecto reducido, las diferencias entre grupos variaron según la tarea, alcanzando su mayor magnitud en T4 y la menor en T2.

Los análisis replicados en la submuestra equilibrada ($n = 1249$ por grupo) mostraron un patrón general consistente con el observado en la muestra completa: se identificaron efectos principales significativos de la ortografía, $F(1, 2496) = 213.47, p < .001$, y del tipo de tarea, $F(3, 7488) = 1320.58, p < .001$, así como una interacción ortografía $\times$ tarea, $F(3, 7488) = 19.42, p < .001$. Los tamaños del efecto fueron comparables, e incluso mayores en algunos casos, a los del análisis principal ($\eta^2p \approx .08$ para ortografía, $\eta^2p \approx .35$ para tarea y $\eta^2p \approx .008$ para la interacción). Este patrón confirma que los resultados observados no dependen del desbalance entre grupos y que se mantienen bajo condiciones de mayor equilibrio muestral.

Gráfico 1

Medias marginales por tarea en función de la ortografía



Nota. Las barras de error representan ± 1 error estándar de las medias marginales estimadas.

3.3 Comparaciones por familia lingüística

El efecto de la familia lingüística, $F(6, 8694) = 108.08, p < .001, \eta^2p \approx .07$, sobre los resultados observados fue significativo; también se observó un efecto significativo del tipo de tarea, $F(3, 26082) = 260.35, p < .001, \eta^2p \approx .03$. La interacción entre familia lingüística y tarea fue significativa, $F(18, 26082) = 2.87, p < .001$, aunque su tamaño del efecto fue reducido ($\eta^2p < .01$).

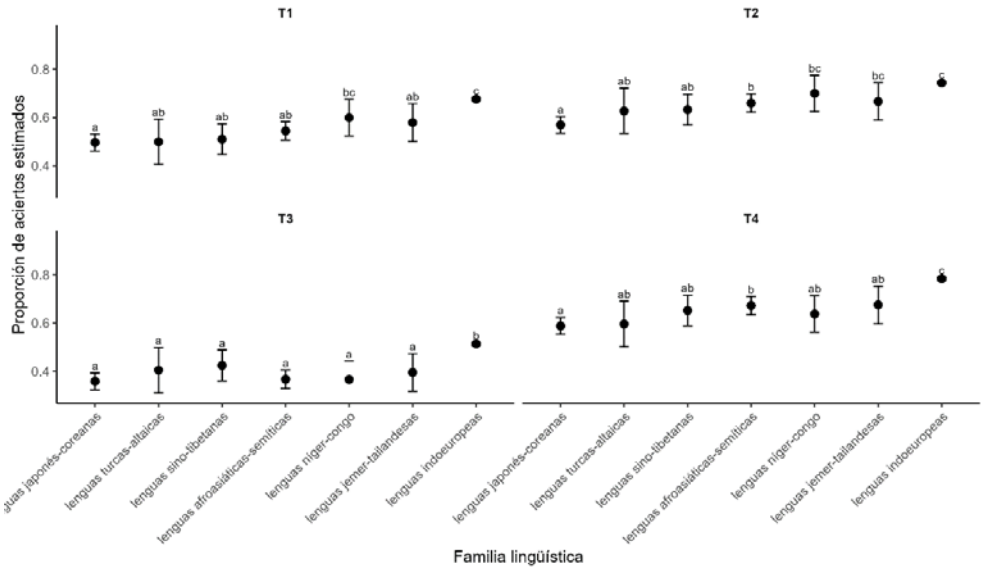
En T1, las medias oscilaron entre 0.497 y 0.676; en T2, entre 0.569 y 0.743; en T3, entre 0.359 y 0.514; y en T4, entre 0.589 y 0.784. La familia indoeuropea presentó las medias más altas, mientras que la familia japonés-coreana obtuvo, en promedio, valores menores; la dispersión de las medias entre familias fue menor en T3 y mayor en T1 y T4 (Gráfico 2).

Las comparaciones post hoc con ajuste de Tukey reflejaron contrastes significativos entre familias lingüísticas en las cuatro tareas, sobre un total de 21 comparaciones pareadas posibles por tarea. Los tamaños del efecto promedio oscilaron entre $d \approx 0.37$ en T3 y $d \approx 0.48$ en T1, con valores similares en T2 y T4 ($d \approx 0.42-0.47$). Las diferencias de mayor magnitud se observaron sistemáticamente entre las lenguas indoeuropeas y las japonés-coreanas, con tamaños del efecto entre $d = 0.98$ y $d = 1.23$.

El Gráfico 2 presenta las medias marginales estimadas para las siete familias lingüísticas en cada tarea, junto con las agrupaciones derivadas de las comparaciones post hoc con ajuste de Tukey. T3 mostró una diferenciación más clara entre familias lingüísticas, mientras que T1, T2 y T4 presentaron patrones más graduales y con mayor superposición entre agrupaciones. En conjunto, los resultados evidencian que las diferencias entre familias lingüísticas variaron según el tipo de tarea.

Gráfico 2

Medias marginales estimadas por familia lingüística y tarea (agrupaciones de Tukey, CLD)



Nota. Las barras de error representan ± 1 error estándar de las medias marginales estimadas.

En la submuestra equilibrada, el efecto de la familia lingüística se mantuvo, $F(6, 196) = 4.67, p < .001$; al igual que para el tipo de tarea $F(3, 588) = 96.10, p < .001$. La interacción entre familia lingüística y tarea no alcanzó la significación estadística, $F(18, 588) = 1.57, p = .061$. Las diferencias entre familias mantuvieron el mismo sentido al observado en la muestra completa.

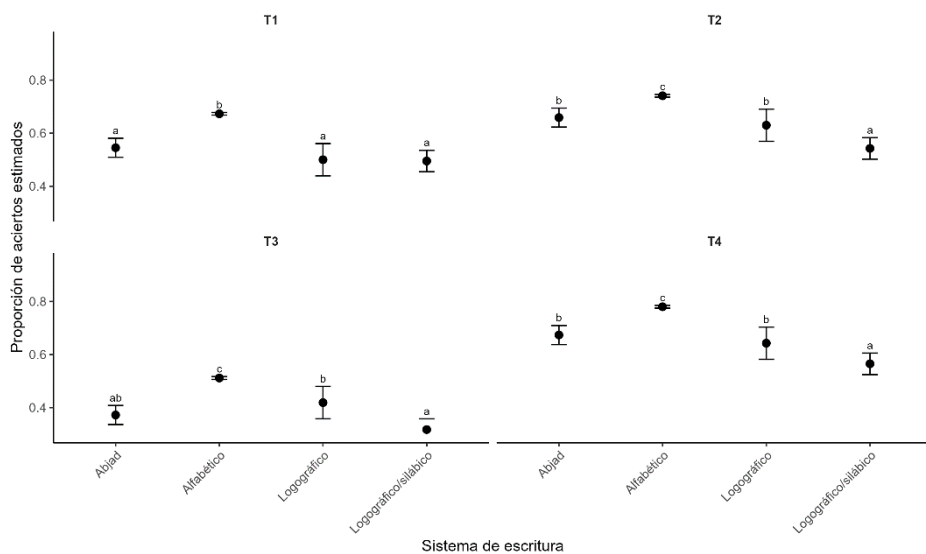
3.4 Comparaciones por sistema de escritura

El modelo lineal mixto mostró un efecto principal significativo del sistema de escritura, $F(3, 8695) = 159.88, p < .001$, con un tamaño de efecto estimado de $\eta^2p \approx .05$. También se observó un efecto significativo del tipo de tarea, $F(3, 26085) = 297.57, p < .001$, con $\eta^2p \approx .03$. La interacción entre sistema de escritura y tarea fue significativa $F(9, 26085) = 2.82, p = .003$ con un tamaño de efecto pequeño ($\eta^2p \approx .001$).

En T1, las medias marginales estimadas oscilaron entre 0.495 (logográfico/silábico) y 0.673 (alfabético). En T2, las puntuaciones medias se situaron entre 0.543 y 0.741. En T3, los valores fueron, en general, más bajos, con medias entre 0.317 y 0.512. En T4, las puntuaciones oscilaron entre 0.565 y 0.780. En las cuatro tareas, el sistema alfabético obtuvo las medias más altas, mientras que el sistema logográfico/silábico mostró las más bajas (Gráfico 3).

Gráfico 3

Medias marginales estimadas por sistema de escritura y tarea (agrupaciones de Tukey, CLD)



Nota. Las barras de error representan ± 1 error estándar de las medias marginales estimadas.

Las comparaciones post hoc con ajuste de Tukey evaluaron seis contrastes posibles por tarea, de los cuales tres resultaron significativos en T1, cuatro en T3 y cinco en T2 y T4. Los tamaños del efecto promedio fueron de magnitud moderada a alta (T1: $d = 0.61$; T2: $d = 0.66$; T3: $d = 0.66$; T4: $d = 0.71$). Los contrastes de mayor magnitud correspondieron sistemáticamente a la comparación entre los sistemas alfabéticos y los logográficos/silábicos, con tamaños del efecto entre $d = 1.12$ y $d = 1.36$. En términos

de rendimiento, estas diferencias equivalen aproximadamente a 1 ítem en T1 (6 ítems), 2 ítems en T2 (10 ítems), 1.5 ítems en T3 (8 ítems) y más de 3 ítems en T4 (14 ítems). La representación gráfica de las diferencias se resume en la figura 3.

Los análisis sobre la submuestra equilibrada reprodujeron el mismo patrón: efectos principales e interacción significativos ($F(3, 236) = 16.89, p < .001$; $F(3, 708) = 134.05, p < .001$; $F(9, 708) = 3.87, p < .001$).

4. Discusión

Medir la comprensión lectora en L2 no consiste únicamente en asignar puntuaciones, sino en garantizar que estas respalden inferencias válidas sobre el constructo evaluado (Elosua, 2024). En contextos de alto impacto, como los exámenes de certificación, la comparabilidad de las puntuaciones puede verse condicionada tanto por las características lingüísticas de los examinados como por el formato de las tareas utilizadas en la evaluación. En este marco, el presente estudio analizó en qué medida el formato del ítem, la transparencia ortográfica, el sistema de escritura y la familia lingüística se asocian con diferencias en el desempeño. En conjunto, los resultados muestran que las diferencias asociadas al formato de la tarea son más consistentes que las relacionadas con las variables lingüísticas analizadas cuyos efectos varían entre tareas.

Respecto a los diferentes formatos de ítems evaluados, las tareas tipo *cloze* selectivo, tendieron a asociarse con menores niveles de desempeño promedio, mientras que opción múltiple y correspondencias mostraron resultados relativamente mayores. Este patrón coincide con investigaciones previas que sugieren que los distintos formatos pueden involucrar demandas de lectura y estrategias de resolución diferentes (Ozuru et al., 2013; Brunfaut & McCray, 2015; Lim, 2019).

En relación con la transparencia ortográfica, los hablantes de lenguas transparentes tendieron a obtener mejores resultados que aquellos provenientes de ortografías opacas. Este resultado es consistente con estudios que han vinculado la consistencia grafema-fonema con diferencias en el desarrollo y automatización de habilidades lectoras (Pae, 2018; Landerl et al., 2019; Shepperd, 2024; Zhang et al., 2025). En este sentido, la mayor regularidad ortográfica podría favorecer procesos de reconocimiento más eficientes durante la lectura.

De manera similar, los examinados asociados a sistemas de escritura alfabéticos mostraron, en promedio, un mejor desempeño que aquellos vinculados a sistemas logográficos o silábicos. Algunas comparaciones específicas alcanzaron tamaños del efecto elevados, especialmente entre los grupos tipológicamente más distantes. Estos resultados encuentran apoyo en investigaciones que señalan que las propiedades estructurales de los sistemas de escritura —incluyendo la complejidad gráfica y las demandas asociadas al reconocimiento léxico— pueden relacionarse con variaciones

en el desempeño lector en L2 (Inoue et al., 2017; Verhoeven & Perfetti, 2022; Hemelstrand & Inoue, 2024).

Respecto a la distancia lingüística, los resultados sugieren que los grupos cuyas lenguas presentan mayor proximidad tipológica al español obtuvieron, en general, mejores resultados que aquellos con lenguas más distantes, en ese sentido los hablantes de lenguas indoeuropeas tendieron a mostrar puntuaciones más altas, mientras que los grupos más distantes —como las lenguas japonés-coreanas— tuvieron, en promedio, valores más bajos. Este patrón converge con estudios que señalan que la proximidad entre la L1 y la L2 puede favorecer el desempeño lector, en la medida en que facilita procesos de transferencia lingüística positiva vinculados al reconocimiento léxico, la integración sintáctica y el aprovechamiento de regularidades compartidas entre lenguas (Jeon & Yamashita, 2014; Zhang & Zhang, 2022; Chai & Bao, 2023).

Aunque los tamaños del efecto de las interacciones fueron reducidos, los resultados mostraron que las diferencias asociadas a las variables lingüísticas no se distribuyeron de manera completamente uniforme entre formatos. En el caso de la transparencia ortográfica, la ventaja del grupo con ortografía transparente fue más marcada en T4 (cloze selectivo), tanto en número de ítems correctamente respondidos como en tamaño del efecto. Del mismo modo, las diferencias entre sistemas de escritura tendieron a ampliarse en las tareas tipo cloze, particularmente en T4, mientras que en formatos como la opción múltiple (T1) las distancias fueron menores. En el caso de las familias lingüísticas, las diferencias variaron según el tipo de tarea, mostrando patrones más graduales en algunos formatos y agrupaciones más claramente diferenciadas en otros. En resumen, estos datos sugieren que, aunque la dirección general de las diferencias entre grupos lingüísticos fue relativamente estable, su expresión concreta varió parcialmente según el formato de la tarea.

La literatura especializada ha señalado que el formato del ítem puede influir en las demandas cognitivas asociadas a las tareas, relacionando los formatos tipo cloze con mayores requerimientos de integración léxica, sintáctica y semántica y los ítems de opción múltiple con estrategias de reconocimiento y localización de información (Brown, 2013; Kobayashi, 2002a, 2002b; Lim, 2014; McCray & Brunfaut, 2018; Ozuru et al., 2013). En este contexto, las mayores diferencias observadas entre grupos en las tareas tipo cloze sugieren que este formato se relacionó de manera diferencial con el desempeño según las características lingüísticas de los examinados. Este patrón fue especialmente visible en los sistemas de escritura, donde las distancias entre los sistemas alfabéticos y los logográficos/silábicos tendieron a ampliarse en este tipo de tareas. No obstante, dado que el presente estudio se basa exclusivamente en medidas de rendimiento, estos resultados no permiten identificar de manera directa los procesos cognitivos implicados, sino únicamente describir como varía el desempeño entre tareas y grupos lingüísticos.

Desde la perspectiva de la validez, los resultados refuerzan la relevancia del formato del ítem en la interpretación de las puntuaciones, más allá de su función operativa en el diseño de la prueba. En línea con los enfoques unificados de validez (Messick, 1989, 1995) y los modelos que enfatizan la interacción entre habilidades y condiciones de tarea (Bachman & Palmer, 1996), los resultados muestran que las diferencias de rendimiento entre grupos lingüísticos no se expresan de manera uniforme en todos los formatos de tarea. En consecuencia, estos resultados respaldan la necesidad de enfoques de validación que consideren no solo el constructo evaluado, sino también las características de las tareas y sus posibles implicaciones para la comparabilidad, la equidad y la interpretación de las puntuaciones (Bachman & Palmer, 1996, 2010; Fulcher, 2015).

En conjunto, los resultados refuerzan la necesidad de seguir examinando la comparabilidad de las puntuaciones en contextos multilingües. En esta línea, futuras investigaciones podrían incorporar análisis de funcionamiento diferencial del ítem (DIF) para determinar si determinadas tareas o formatos favorecen o desfavorecen sistemáticamente a grupos con distintos perfiles lingüísticos (Elosua, 2024). Asimismo, analizar el comportamiento de distintos formatos en grupos lingüísticos diversos y pilotar tareas en muestras con diferentes perfiles de L1 podría aportar evidencia adicional sobre qué configuraciones resultan más sensibles a diferencias tipológicas entre examinados.

CONCLUSIONES

La investigación previa ha documentado por separado la influencia de la L1 en la lectura en L2 (Frost, 2012; Jeon & Yamashita, 2014) y el impacto del formato del ítem en el rendimiento (Bachman & Palmer, 1996; Liao, 2023); sin embargo, pocos estudios han examinado su interacción. En este sentido, los resultados obtenidos muestran que las diferencias asociadas a las características lingüísticas de la L1 variaron según el formato de la tarea, lo que refuerza la relevancia de considerar conjuntamente ambos elementos en la interpretación del desempeño en comprensión lectora. Esta conclusión conecta con estudios sobre procesos (Brunfaut & McCray, 2015; McCray & Brunfaut, 2018) y amplía la evidencia al plano del rendimiento observable en evaluaciones de gran escala.

A pesar de la solidez metodológica del estudio, los resultados deben interpretarse considerando algunas limitaciones, especialmente el uso del país de aplicación como aproximación indirecta a la lengua materna, la ausencia de variables individuales y el análisis de un único nivel de competencia del MCER. Futuras investigaciones podrían incorporar información más precisa sobre el perfil lingüístico de los participantes — como la lengua materna autoinformada, el grado de exposición al español o la trayectoria de aprendizaje—, así como examinar distintos niveles de competencia MCER y contextos de evaluación.

Dado que el estudio se basó exclusivamente en medidas de rendimiento observable, no fue posible examinar de manera directa los procesos implicados en la resolución de las tareas. Futuras investigaciones podrían incorporar medidas de proceso, como eye-tracking o protocolos de recuerdo estimulado, además de analizar versiones alternativas de tareas con diferentes demandas contextuales u ortográficas.

Estos hallazgos contribuyen a una comprensión más matizada de la evaluación de la comprensión lectora en L2 y aporta evidencia relevante para avanzar hacia la construcción de instrumentos de evaluación más comparables entre grupos lingüísticos, en coherencia con enfoques contemporáneos sobre validez y evaluación lingüística.

REFERENCIAS BIBLIOGRÁFICAS

- Alderson, J. C. (2000). *Assessing reading*. Cambridge University Press.
- American Educational Research Association, American Psychological Association, & National Council on Measurement in Education. (2014). *Standards for educational and psychological testing*. American Educational Research Association.
- Association of Language Testers in Europe. (2024). Setting standards: The ALTE Q-mark. ALTE.
- Baayen, R. H., Davidson, D. J., & Bates, D. M. (2008). Mixed-effects modeling with crossed random effects for subjects and items. *Journal of Memory and Language*, 59(4), 390-412.
- Bachman, L. F., & Palmer, A. S. (1996). *Language testing in practice: Designing and developing useful language tests*. Oxford University Press.
- Bachman, L. F., & Palmer, A. S. (2010). *Language assessment in practice: Developing language assessments and justifying their use in the real world*. Oxford University Press.
- Balota, D. A., Flores d'Arcais, G. B., & Rayner, K. (1990). *Comprehension processes in reading*. Lawrence Erlbaum Associates.
- Becker, A., & Nekrasova-Beker, T. (2018). Investigating the effect of different selected-response item formats for reading comprehension. *Educational Assessment*, 23(4), 296-317.
- Bobb, S. C., Von Holzen, K., Mayor, J., Mani, N., & Carreiras, M. (2020). Co-activation of the L2 during L1 auditory processing: An ERP cross-modal priming study. *Brain and Language*, 203, 104739.
- Brown, J. D. (2013). My twenty-five years of cloze testing research: So what. *International Journal of Language Studies*, 7(1), 1-32.

- Brunfaut, T., & McCray, G. (2015). Looking into test-takers' cognitive processes while completing reading tasks: A mixed-method eye-tracking and stimulated recall study. *British Council/UALTA Research Papers*, 2015(1), 1-65.
- Cambridge University Press & Assessment. (2022). *B2 First: Reading and Use of English – Sample Test 1*. Cambridge University Press & Assessment.
- Cargnelutti, E., Tomasino, B., & Fabbro, F. (2022). Effects of linguistic distance on second language brain activations in bilinguals: An exploratory coordinate-based meta-analysis. *Frontiers in Human Neuroscience*, 15, 744489.
- Chai, X., & Bao, J. (2023). Linguistic distances between native languages and Chinese influence acquisition of Chinese character, vocabulary, and grammar. *Frontiers in Psychology*, 13, 1083574.
- Cohen, J. (1988). *Statistical Power Analysis for the Behavioral Sciences* (2nd ed.). Lawrence Erlbaum Associates.
- Consejo de Europa. (2002). *Marco común europeo de referencia para las lenguas: aprendizaje, enseñanza, evaluación*. Ministerio de Educación, Cultura y Deporte / Instituto Cervantes.
- Cummins, J. (1979). Linguistic interdependence and the educational development of bilingual children. *Review of Educational Research*, 49, 222-251.
- Cummins, J. (1981). The role of primary language development in promoting educational. *Schooling and language minority students: A theoretical framework*, 16.
- Elosua, P. (2024). A three-step DIF analysis of a reading comprehension test across regional dialects to improve test score validity. *Language Assessment Quarterly*, 21(2), 141-158.
- Frost, R. (2012). A universal approach to modeling visual word recognition and reading: Not only possible, but also inevitable. *The behavioral and brain sciences*, 35(5), 310.
- Fulcher, G. (2015). *Re-examining Language Testing: A Philosophical and Social Inquiry*. Routledge.
- Gernsbacher, M. A. (1990). *Language comprehension as structure building*. Lawrence Erlbaum Associates.
- Goethe-Institut. (2018). *Goethe-Zertifikat B2: Prüfungsordnung, Durchführungsbestimmungen, Hinweise*. Goethe-Institut.
- Grabe, W. (2009). *Reading in a second language: Moving from theory to practice*. Cambridge University Press.

- Graesser, A. C. (2015). Deeper learning with advances in discourse science and technology. *Policy Insights from the Behavioral and Brain Sciences*, 2(1), 42-50.
- Graesser, A. C., Singer, M., & Trabasso, T. (1994). Constructing inferences during narrative text comprehension. *Psychological Review*, 101, 371-395.
- Haladyna, T. M., & Rodriguez, M. C. (2013). *Developing and validating test items* (1st ed.). Routledge.
- Hao, Y., Luo, Y., Lin-Hong, K. H. Y., & Yan, M. (2024). Shared translation in second language activates unrelated words in first language. *Psychonomic Bulletin & Review*, 31(3), 1245-1255.
- Hemelstrand, S., & Inoue, T. (2024). The extrema of Japanese Literacy: beyond the bounds of average reading. *Reading Research Quarterly*, 59(3), 313-326.
- Hu, L., & Bentler, P. M. (1999). Cutoff criteria for fit indexes in covariance structure analysis: Conventional criteria versus new alternatives. *Structural Equation Modeling: A Multidisciplinary Journal*, 6(1), 1-55.
- Inoue, T., Georgiou, G. K., Muroya, N., Maekawa, H., & Parrila, R. (2017). Cognitive predictors of literacy acquisition in syllabic Hiragana and morphographic Kanji. *Reading and Writing*, 30, 1335-1360.
- Instituto Cervantes. (2012). *DELE B2: Examen 00, versión 1 (septiembre 2012)*. Instituto Cervantes. <https://exámenes.cervantes.es/es/dele/exámenes/b2>
- Jeon, E. H., & Yamashita, J. (2014). L2 reading comprehension and its correlates: A meta-analysis. *Language Learning*, 64, 160-212.
- Jonz, J. (1976). Improving on the basic egg: The m-c cloze. *Language Learning*, 26(2), 255-265.
- Kane, M. (2013). Validating the interpretations and uses of test scores. *Journal of Educational Measurement*, 50(1), 1-73.
- Katz, L., & Frost, R. (1992). The reading process is different for different orthographies: The orthographic depth hypothesis. *Advances in Psychology*, 94, 67-84.
- Kintsch, W. (1988). The role of knowledge in discourse comprehension: A construction–integration model. *Psychological Review*, 95, 163-182.
- Kobayashi, M. (2002a). Cloze tests revisited: Exploring item characteristics with special attention to scoring methods. *The Modern Language Journal*, 86(4), 571-586.

- Kobayashi, M. (2002b). Method effects on reading comprehension test performance: Text organization and response format. *Language Testing*, 19(2), 193-220.
- Landerl, K., Freudenthaler, H. H., Heene, M., de Jong, P., Desroches, A., Manolitsis, G., . . . Georgiou, G. K. (2019). Phonological awareness and rapid automatized naming as longitudinal predictors of reading in five alphabetic orthographies with varying degrees of consistency. *Scientific Studies of Reading*, 23, 220-234.
- Landerl, K., Castles, A., & Parrila, R. (2022). Cognitive precursors of reading: A cross-linguistic perspective. *Scientific Studies of Reading*, 26(2), 111-124.
- Liao, R. J. (2023). The use of generalizability theory in investigating the score dependability of classroom-based L2 reading assessment. *Language Testing*, 40(1), 86-106.
- Lim, H. J. (2014). *Exploring the validity evidence of the TOEFL IBT reading test from a cognitive perspective*. Michigan State University. Second Language Studies.
- Lim, H. (2019). Test format effects: A componential approach to second language reading. *Language Testing in Asia*, 9(1), 6.
- Luke, S. G. (2017). Evaluating significance in linear mixed-effects models. *Behavior Research Methods*, 49, 1494-1502.
- McBride, C., Pan, D. J., & Mohseni, F. (2022). Reading and writing words: A cross-linguistic perspective. *Scientific Studies of Reading*, 26(2), 125-138.
- McCray, G., & Brunfaut, T. (2018). Investigating the construct measured by banked gap-fill items: Evidence from eye-tracking. *Language Testing*, 35(1), 51-73.
- McDonald, R. P. (2013). *Test theory: A unified treatment*. Psychology Press.
- Messick, S. (1995). Validity of psychological assessment: Validation of inferences from persons' responses and performances as scientific inquiry into score meaning. *American Psychologist*, 50(9), 741-749.
- Messick, S. (1989). Validity. En R. L. Linn (Ed.), *Educational measurement* (pp. 13-103). American Council on Education; Macmillan.
- Muñoz, C., Cadierno, T., & Casas, I. (2018). Different starting points for English language learning: A comparative study of Danish and Spanish young learners. *Language Learning*, 68(4), 1076-1109.
- Myers, J. L., & O'Brien, E. J. (1998). Accessing the discourse representation during reading. *Discourse Processes*, 26, 131-157.

- O'Sullivan, B., & Weir, C. J. (2011). Test development and validation. En B. O'Sullivan (Ed.), *Language testing: Theories and practices* (pp. 13-32). Palgrave Macmillan.
- Oller, J. W., & Jonz, J. (1994). Why cloze procedure? En J. W. Oller & J. Jonz (Eds.), *Cloze and coherence* (pp. 1-20). Bucknell University Press.
- Ozuru, Y., Briner, S., Kurby, C. A., & McNamara, D. S. (2013). Comparing comprehension measured by multiple-choice and open-ended questions. *Canadian Journal of Experimental Psychology/Revue canadienne de psychologie expérimentale*, 67(3), 215-227.
- Pae, H. (2018). *Writing systems, reading processes, and cross-linguistic influences: Reflections from the Chinese, Japanese and Korean languages*. John Benjamins.
- Perfetti, C., & Stafura, J. (2014). Word knowledge in theory of reading comprehension. *Scientific Studies of Reading*, 18(1), 22-37.
- Piepho, H. P. (2004). An Algorithm for a Letter-Based Representation of All-Pairwise Comparisons. *Journal of Computational and Graphical Statistics*, 13(2), 456-466.
- Radman, N., Jost, L., Dorood, S., Mancini, C., & Annoni, J. M. (2021). Language distance modulates cognitive control in bilinguals. *Scientific Reports*, 11(1), 24131.
- Schepens, J. J., Van der Slik, F., & Van Hout, R. (2016). L1 and L2 distance effects in learning L3 Dutch. *Language Learning*, 66(1), 224-256.
- Shepperd, L. (2024). Cross-scriptal orthographic influence on second language phonology. *Languages*, 9(6), 210.
- Sproat, R., & Gutkin, A. (2021). The taxonomy of writing systems: How to measure how logographic a system is. *Computational Linguistics*, 47(3), 477-528.
- Università per Stranieri di Perugia. (2017). *CELI 3 – Certificato di Conoscenza della Lingua Italiana*. Università per Stranieri di Perugia.
- Ureña, C., Flores, M. E., & Martín, S. (2024). Competencia comunicativa oral en español en contextos de migración: diseño y validación de un examen multinivel. *Revista Signos*, 57(115), 594-618.
- van den Broek, P., Young, M., Tzeng, Y., & Linderholm, T. (1999). The landscape model of reading. En H. van Oostendorp & S. R. Goldman (Eds.), *The construction of mental representations during reading* (pp. 71-98). Lawrence Erlbaum Associates.

- Verhoeven, L., & Perfetti, C. (2022). Universals in learning to read across languages and writing systems. *Scientific Studies of Reading*, 26(2), 150-164.
- Wang, Y., Arslan-Ari, I., Hao, L., & Hwang, K. (2024). Reading alphabetic and nonalphabetic writing systems: A case study of bilingual teachers' reading processes through eye movement miscue analysis. *Journal of Adolescent & Adult Literacy*, 68(3), 223-233.
- Weir, C. J. (2005). Language testing and validation. *Hampshire: Palgrave MacMillan*, 10, 9780230514577.
- West, B. T., Welch, K. B., & Galecki, A. T. (2014). *Linear mixed models: A practical guide using statistical software* (2nd ed.) Routledge.
- Yamashita, H. (2008). Effects of sentence processing strategy proximity on the comprehension of second languages. *Second Language*, 7, 43-82.
- Zeigenfuse, M. D., Batchelder, W. H., & Steyvers, M. (2020). An item response theory model of matching test performance. *Journal of Mathematical Psychology*, 95, 102327.
- Zhang, S., & Zhang, X. (2022). The relationship between vocabulary knowledge and L2 reading/listening comprehension: A meta-analysis. *Language Teaching Research*, 26(4), 696-725.
- Zhang, K., Sun, X., Flores-Gaona, Z., Yu, C. L., Eggleston, R. L., Nickerson, N., ... & Kovelman, I. (2025). Cross-linguistic transfer in bilingual children's phonological and morphological awareness skills: a longitudinal perspective. *Bilingualism: Language and Cognition*, 28(2), 327-342.

AGRADECIMIENTOS

Este trabajo ha sido parcialmente financiado por el MICIU/AEI 10.13039/501100011033 (proyecto PID2024-155252OB-I00).