

Demanda cognitiva en las pruebas estandarizadas de Matemáticas en España y Portugal

Laia Tugores, Ana Paula Aires & Manuela Raposo Rivas

Resumen

Las pruebas de evaluación estandarizadas constituyen un valioso instrumento para orientar las políticas educativas. El propósito de esta investigación es cuantificar las demandas cognitivas de las pruebas de Matemáticas realizadas en 2024 y 2025 en la comunidad autónoma de Galicia (España) y Portugal, para compararlas con las demandas curriculares de los cursos correspondientes, con el reparto de Lange y con las pruebas PISA y TIMSS. Para ello, se utilizan los procesos de la taxonomía de Bloom revisada. La metodología es mixta, con un enfoque descriptivo. En Galicia se analizan las pruebas de diagnóstico para alumnado de 9 y 13 años, así como las pruebas de acceso a la Universidad (17 años); en Portugal, las pruebas finales realizadas a los 14, 16 y 17 años. La información de cada prueba se recoge en una ficha de registro en forma de tabla de doble entrada (ejercicios / procesos cognitivos). Los datos obtenidos se analizan estadísticamente. Se obtiene que las demandas cognitivas de estas pruebas son mayores que la demanda asociada al modelo de Lange, pero inferiores a las curriculares, PISA y TIMSS. Las distribuciones de los procesos cognitivos presentan diferencias significativas respecto a PISA y TIMSS. Ampliar este estudio a otros países permitiría diseñar mapas de demanda y elaborar repartos cognitivos ajustados a las distintas realidades nacionales.

Palabras clave:

Demanda cognitiva; Galicia; Matemáticas; Portugal; Pruebas estandarizadas.

Cognitive Demand in Standardized Mathematics Assessments in Spain and Portugal

Abstract: Standardized assessments constitute a valuable tool for informing educational policy. The aim of this study is to quantify the cognitive demands of the Mathematics tests administered in 2024 and 2025 in the autonomous community of Galicia (Spain) and in Portugal, in order to compare them with the curricular requirements for the corresponding year groups, with Lange's distribution, and the PISA and TIMSS assessments. To this end, the cognitive processes of the revised Bloom's taxonomy are used. The methodology is mixed, with a descriptive approach. In Galicia, diagnostic assessments for students aged 9 and 13, as well as the university entrance examinations (age 17), are analysed; in Portugal, the final examinations taken at ages 14, 16, and 17 are examined. Information from each test is recorded in a data collection sheet structured as a double-entry table (items / cognitive processes). The resulting data are analysed statistically. The findings show that the cognitive demands of these tests are higher than those associated with Lange's model, but lower than the curricular demands and those observed in PISA and TIMSS. The distributions of cognitive processes show significant differences with respect to PISA and TIMSS. Extending this study to other countries would make it possible to design demand maps and to develop cognitive distributions better aligned with different national contexts.

Keywords: Cognitive demand; Galicia; Mathematics; Portugal; Standardized assessments.

Demande cognitive dans les épreuves standardisées de mathématiques en Espagne et au Portugal

Résumé: Les épreuves d'évaluation standardisées constituent un instrument précieux pour orienter les politiques éducatives. L'objectif de cette recherche est de quantifier les exigences cognitives des épreuves de mathématiques administrées en 2024 et 2025 dans la communauté autonome de Galice (Espagne) et au Portugal, afin de les comparer aux exigences curriculaires des niveaux correspondants, à la répartition de Lange ainsi qu'aux évaluations PISA et TIMSS. À cette fin, les processus de la taxonomie révisée de Bloom sont utilisés. La méthodologie est mixte, avec une approche descriptive. En Galice, les épreuves diagnostiques destinées aux élèves de 9 et 13 ans sont analysées, ainsi que les épreuves d'accès à l'université (17 ans) ; au Portugal, les épreuves finales passées à 14, 16 et 17 ans sont examinées. Les informations de chaque épreuve sont recueillies dans une fiche d'enregistrement sous la forme d'un tableau à double entrée (exercices / processus cognitifs). Les données obtenues sont analysées statistiquement. Les résultats montrent que les exigences cognitives de ces épreuves sont supérieures à celles associées au modèle de Lange, mais inférieures aux exigences curriculaires ainsi qu'à celles de PISA et TIMSS. Les distributions des processus cognitifs présentent des différences significatives par rapport à PISA et TIMSS. L'extension de cette étude à d'autres pays permettrait de concevoir des cartes de demande et d'élaborer des répartitions cognitives adaptées aux différentes réalités nationales.

Mots-clés : Demande cognitive ; Épreuves standardisées ; Galice ; Mathématiques ; Portugal.

Exigência cognitiva nas provas estandardizadas de Matemática em Espanha e Portugal

Resumo: As provas de avaliação estandardizadas constituem um instrumento valioso para orientar as políticas educativas. O objetivo desta investigação é quantificar as exigências cognitivas das provas de Matemática realizadas em 2024 e 2025 na comunidade autónoma da Galiza (Espanha) e em Portugal, para as comparar com as exigências curriculares dos anos de escolaridade correspondentes, com a distribuição de Lange e com as provas PISA e TIMSS. Para tal, utilizam-se os processos da taxonomia revista de Bloom. A metodologia é mista, com uma abordagem descritiva. Na Galiza analisam-se as provas de diagnóstico para alunos de 9 e 13 anos, bem como as provas de acesso à universidade (17 anos); em Portugal, as provas finais realizadas aos 14, 16 e 17 anos. A informação de cada prova é registada numa ficha de registo sob a forma de uma tabela de dupla entrada (exercícios / processos cognitivos). Os dados obtidos são analisados estatisticamente. Os resultados mostram que as exigências cognitivas destas provas são superiores às associadas ao modelo de Lange, mas inferiores às exigências curriculares e às de PISA e TIMSS. As distribuições dos processos cognitivos apresentam diferenças significativas em relação a PISA e TIMSS. A extensão deste estudo a outros países permitiria conceber mapas de exigência e elaborar distribuições cognitivas ajustadas às diferentes realidades nacionais.

Palavras-chave: Demanda cognitiva; Galiza; Matemática; Portugal; Provas de avaliação estandardizadas.

Introducción

Las pruebas estandarizadas son evaluaciones de tipo censal que se llevan a cabo en momentos concretos del proceso educativo para medir niveles competenciales del alumnado. Gutiérrez y Acuña (2022) elaboran una revisión sistemática de estos instrumentos y destacan el uso de marcos de referencia teóricos y metodológicos sólidos, su calidad técnica y su rigor científico; no obstante, advierten del riesgo de emplearlas para valorar aspectos educativos que exceden su propósito. Además, focalizar la enseñanza en la preparación sistemática de estas pruebas, con el objetivo de mejorar posiciones en *rankings* de calidad, puede alejar a los estudiantes de un aprendizaje necesariamente neutro.

Auld y Morris (2013) señalan la división existente entre los investigadores respecto a la utilidad de las pruebas estandarizadas en la educación comparada, por su orientación explícita a las políticas educativas y su dependencia de indicadores estadísticos de logro. Numerosos autores consideran que no es posible establecer conexiones claras de causa y efecto que justifiquen medidas correctoras en los sistemas de enseñanza, dado el limitado conocimiento del contexto en que estos se insertan y la influencia determinante de factores históricos y culturales.

Entre sus defensores, Shepard (2006) subraya que las pruebas estandarizadas evalúan lo que es más importante aprender, a diferencia de las tradicionales, centradas en lo más fácil de medir. Entre sus detractores, Popham (1999) recurre a la exageración al afirmar que “emplear pruebas estandarizadas de logros para averiguar la calidad educativa es como medir la temperatura con una cuchara” (p. 10). En una posición intermedia, Martínez (2009) aboga por sistemas de evaluación que combinen de forma más equilibrada las pruebas externas y las prácticas evaluativas del aula.

Por su parte, Fernández *et al.* (2017) sostienen que la complejidad de las competencias dificulta medirlas con precisión y que estas pruebas masivas ni siquiera pertenecen propiamente al campo de la evaluación; aun así, reconocen que, bajo un modelo estímulo–respuesta, pueden aportar información útil sobre el proceso formativo. Asimismo, Da Silva *et al.* (2020) advierten que la evaluación del sistema educativo no debe limitarse a las calificaciones obtenidas en pruebas estandarizadas y reclaman un enfoque sistémico que incorpore factores como la infraestructura escolar, el contexto socioeconómico y cultural del alumnado y la formación docente continua.

Rodríguez *et al.* (2025) destacan que las evaluaciones estandarizadas han avanzado notablemente gracias a la teoría de respuesta al ítem, las pruebas adaptativas y el uso de inteligencia artificial, lo que ha mejorado la precisión de las mediciones; sin

embargo, subrayan que su impacto se ve limitado por la falta de conexión entre diseño y práctica, la escasa formación docente para interpretarlas y el insuficiente aprovechamiento de los datos generados.

Así, Fernandes (2007) observa que los resultados en las pruebas externas suelen ser modestos, especialmente cuando se exige al alumnado movilizar o aplicar conocimientos de forma integrada, aunque no aporta explicaciones sobre las causas de este rendimiento.

En definitiva, pese a su amplia implantación, las pruebas estandarizadas suscitan controversias en torno a su propósito, utilidad y capacidad para reflejar la complejidad del aprendizaje. Ante estas limitaciones y divergencias, resulta necesario analizar las pruebas estandarizadas con una perspectiva que permita valorar la demanda de pensamiento implicada en su resolución.

En la presente investigación se examinan las pruebas estandarizadas desde una óptica exclusivamente cognitiva y se prescinde no solo de las calificaciones, sino también de los contenidos y las idoneidades. Para ello, se considera la dimensión cognitiva estructurada jerárquicamente por Anderson y Krathwohl (2001) en seis procesos, de menor a mayor complejidad: *recordar* (recuperar información previamente aprendida), *comprender* (interpretar significado), *aplicar* (usar conocimientos o procedimientos en situaciones conocidas o análogas), *analizar* (descomponer para identificar relaciones o principios subyacentes), *evaluar* (emitir juicios fundamentados) y *crear* (integrar elementos para producir algo nuevo). Esta revisión de la taxonomía de Bloom, del año 1956, es conocida como RBT (*Revised Bloom's Taxonomy*).

Este modelo RBT para el diseño de pruebas de evaluación fue validado por Pizà-Mir (2022) mediante una muestra integrada por jueces externos, docentes y estudiantes, quienes lo consideraron eficaz para tal cometido, con resultados estadísticos que respaldan su fiabilidad. Asimismo, verifica que su uso promueve el aprendizaje.

Además de su natural vinculación con las pruebas de evaluación, la exigencia cognitiva se ha analizado en otros eslabones del proceso educativo, como las pautas para valorar el aprendizaje (Leite & Valente, 2024), las competencias clave de las asignaturas (Sarmiento & Sarmiento, 2023) y los criterios curriculares de evaluación (Tugores, 2024; Tugores *et al.*, 2025).

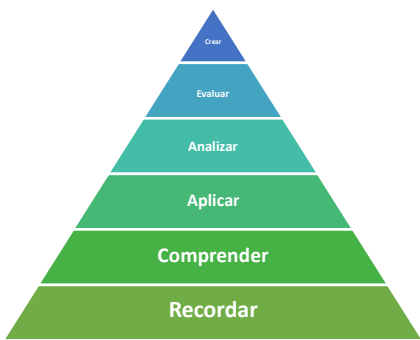
Marco teórico

En este trabajo se considera el modelo cognitivo de Lange (Boertien & Lange, 1994) por su relevancia al integrar cognición, emoción y acción en un sistema único. Bajo la

premisa de que el pensamiento humano no se construye desde lo más complejo hacia lo simple, sino al revés, Lange establece cómo deben distribuirse los procesos cognitivos de una estructura jerárquica con n niveles de complejidad en una prueba de evaluación. En un triángulo equilátero, se trazan $n-1$ segmentos paralelos a uno de sus lados y equidistantes entre sí, generando así n regiones. A cada nivel se le asigna una región, de modo que, cuanto mayor es la complejidad del nivel, menor es el área correspondiente.

Según Lange, una prueba está “cognitivamente equilibrada” cuando la frecuencia relativa de aparición de cada nivel coincide con la proporción del área de su región. A esta distribución cognitiva se le denomina *reparto Lange*. El triángulo asociado a la taxonomía RBT se muestra en la Figura 1.

Figura 1
Reparto Lange para los procesos cognitivos RBT



Fuente: *Elaboración propia.*

La Tabla 1 presenta los porcentajes de las áreas de las regiones en el reparto Lange, para la RBT ($n = 6$), así como para $n = 4$ y $n = 3$, valores que también se emplean habitualmente.

Tabla 1
Porcentajes de las áreas de las regiones en repartos Lange

n	Porcentajes Menor complejidad → Mayor					
6	30.6	25	19.4	13.9	8.3	2.8
4	43.8	31.2	18.8	6.2		
3	55.6	33.3	11.1			

Fuente: *Elaboración propia.*

Con el fin de evaluar las exigencias cognitivas de las pruebas sin que medien subjetividades, utilizamos una escala denominada *gradación uniforme* (introducida por Tugores (2024) para estimar la demanda cognitiva en los criterios de evaluación de Matemáticas en Educación Secundaria). En ella, cada nivel de complejidad recibe una *valoración cognitiva* igual al numeral de su orden inverso en la jerarquía, de ahí su nombre.

Llamamos *demanda cognitiva* de una prueba, y la denotamos por D , al resultado de sumar las valoraciones cognitivas de los niveles multiplicadas por sus proporciones en la prueba. Esta variable permite comparar exigencias cognitivas de dos pruebas, incluso cuando se han utilizado taxonomías distintas.

Para facilitar la comparación de demandas cognitivas calculadas a partir de taxonomías con distinto número de niveles, todas se trasladan a una escala de 0 a 10. Por ejemplo, algunas demandas cognitivas para los repartos Lange con n niveles, que denotamos por $DL(n)$, son: $DL(3) = 2.78$, $DL(4) = 2.92$ y $DL(6) = 3.06$; obsérvese que todas ellas son sensiblemente menores que la media de la escala (5).

Objetivo del estudio

El objetivo de esta investigación es comparar las demandas cognitivas de las pruebas estandarizadas de Matemáticas realizadas en Galicia y Portugal en 2024 y 2025 con las exigidas en los currículos académicos, el modelo Lange, PISA, TIMSS y entre ellas mismas, empleando la taxonomía RBT. Asimismo, se contrastan las distribuciones de los procesos cognitivos presentes en dichas pruebas. De esta manera, se complementan los estudios previos sobre pruebas estandarizadas, habitualmente centrados en aspectos matemáticos y de logro académico.

Marco contextual

En España, cada comunidad autónoma elabora sus propias pruebas. En este estudio se ha escogido Galicia por las afinidades culturales que mantiene con Portugal, lo que facilita la comparación entre ambos sistemas educativos y el análisis de posibles paralelismos en la concepción y diseño de las evaluaciones externas de matemáticas.

En dicho país se realizan evaluaciones de diagnóstico de la competencia matemática para todo el alumnado a los 9 y 13 años. Sus resultados tienen carácter informativo y se utilizan para orientar a los distintos estamentos educativos. El documento “Marco general de las evaluaciones del sistema educativo” (Ministerio de Educación, Formación Profesional y Deportes, 2026), abreviadamente MGESE 2026, precisa los bloques competenciales, sentidos matemáticos y contextos que se consideran. Dicho documento asigna un grado de dificultad matemática a las preguntas y establece

su distribución en las pruebas, pero no valora complejidades cognitivas. Asimismo, el alumnado de 17 años que desea acceder a la universidad realiza pruebas específicas de acceso.

Caraballo *et al.* (2011) examinaron la complejidad cognitiva de las pruebas de diagnóstico para alumnado de 13 años en diferentes comunidades autónomas españolas (curso 2008-09) mediante las categorías *reproducción*, *conexión* y *reflexión* (orden creciente de complejidad), evidenciando una adecuación parcial del diseño de las pruebas a los niveles de competencia matemática que se pretendía evaluar.

En Portugal, el IAVE (*Instituto de Avaliação Educativa*) organiza las pruebas finales de Matemáticas que el alumnado realiza a los 14, 16 y 17 años (estas dos últimas con influencia en el acceso a la enseñanza universitaria). El IAVE clasifica los ítems en tres categorías cognitivas, de menor a mayor complejidad: *conocer/reproducir*, *aplicar/interpretar* y *razonar/crear*. En 2022 publicó su distribución en la prueba de 14 años de 2021: 40%, 35% y 25%, respectivamente (Instituto de Avaliação Educativa, I. P., 2022).

Según Marques *et al.* (2015), el profesorado portugués apoya claramente la evaluación externa en la Enseñanza Básica mediante exámenes finales de ciclo de ámbito nacional, y muchos docentes profundizan en los contenidos habituales de dichas pruebas y adoptan su lenguaje en el aula.

En relación con las pruebas a los 17 años, Boal *et al.* (2008) sostienen que el alumnado es entrenado para memorizar conceptos y repetir procedimientos matemáticos de forma rutinaria, resolviendo ejercicios típicos una y otra vez en un contexto de aprendizaje superficial, lo que intensifica los procesos RBT *recordar* y *aplicar*. Por el contrario, López *et al.* (2020) afirman que, a esa edad, se priorizan los procesos de mayor complejidad cognitiva.

Pruebas estandarizadas de matemáticas en el ámbito internacional

En Estados Unidos, la Evaluación Nacional del Progreso Educativo (*National Assessment of Educational Progress*, NAEP) aplica en Matemáticas pruebas a muestras representativas de alumnado de 9, 13 y 17 años (*National Assessment Governing Board*, 2026), en las que la demanda cognitiva se organiza en tres niveles (bajo, moderado y alto), pero no se establecen porcentajes por nivel, ya que la complejidad se utiliza para equilibrar el conjunto de ítems.

En educación comparada, Azimi *et al.* (2025) exploran el nivel de demanda cognitiva de los ítems de Matemáticas en exámenes nacionales de acceso a la universidad de Irán, Turquía y Estados Unidos, utilizando el marco *Cognitive Demand Level*, que distingue entre los niveles de menor a mayor complejidad: *memorización*,

procedimientos sin conexiones, procedimientos con conexiones y haciendo Matemáticas. Su estudio muestra un claro predominio de las tareas procedimentales, mientras que los ítems de mayor exigencia conceptual o de razonamiento profundo son minoritarios, lo que se traduce en pruebas con escaso peso de la resolución de ejercicios no rutinarios.

Las evaluaciones estandarizadas más extendidas son PISA (*Programme for International Student Assessment*) y las pruebas TIMSS (*Trends in International Mathematics and Science Study*).

Las pruebas PISA se llevan a cabo a nivel mundial para medir el rendimiento académico de los estudiantes de 15 años en Matemáticas, Ciencias y Lectura. Rosales *et al.* (2020) constatan en las pruebas liberadas de Biología y Geología, entre 2000 y 2015, un predominio de ejercicios con baja demanda cognitiva RBT. Estas pruebas suelen plantear situaciones relacionadas con la vida de los estudiantes, donde el uso de la ciencia sirve de base para plantear nuevas ideas, por lo que su escasa exigencia podría verse compensada por el esfuerzo que requiere analizar los encabezamientos contextualizados.

Las pruebas de Matemáticas PISA 2022 (Ministerio de Educación, Formación Profesional y Deportes, 2023) consideraron cuatro procesos cognitivos, de menor a mayor complejidad: *formulación* de situaciones, *empleo* de conceptos, hechos y procedimientos, *interpretación* y *evaluación* de resultados y *razonamiento* (deductivo e inductivo). En las unidades liberadas, los porcentajes fueron: *formulación* = *empleo* = 20%, *interpretación* y *evaluación* = 35% y *razonamiento* = 25%, con $D = 5.50$, por encima de $DL(4)$. Las pruebas PISA 2025 distinguieron estos mismos procesos.

Por otro lado, las pruebas TIMSS son evaluaciones internacionales de Matemáticas y Ciencias que se realizan a los 9 y 13 años. Los marcos de la evaluación de Matemáticas en el año 2023 contemplaron los dominios cognitivos, de menor a mayor complejidad: *conocimiento*, *aplicación* y *razonamiento*, así como los porcentajes para cada uno (Instituto Nacional de Evaluación Educativa, 2022, pp. 20-21). La Tabla 2 recoge estos porcentajes y los valores de demanda cognitiva.

Tabla 2
Porcentajes de los dominios y valores de las demandas cognitivas en TIMSS 2023

	Conocimiento	Aplicación	Razonamiento	D
9 años	40	40	20	4
13 años	35	40	25	4.5

Fuente: Elaboración propia.

Los valores de D son mayores que $DL(3)$ y menores que el de las pruebas PISA 2022, si bien estas fueron realizadas por estudiantes de 15 años.

Metodología

Se utiliza una metodología mixta con predominio cualitativo (Hernández-Sampieri *et al.*, 2014), orientada a identificar los procesos cognitivos en las pruebas estandarizadas y complementada con datos numéricos para cuantificar la demanda cognitiva. El diseño es descriptivo. Se realiza un análisis de contenido de las pruebas (cuyas unidades son las preguntas) y, seguidamente, un tratamiento estadístico de los datos extraídos. El instrumento de recogida es una ficha de registro, elaborada para cada prueba con formato de tabla de doble entrada: las filas corresponden a las preguntas (identificadas por su ordinal) y las columnas, a los procesos RBT. En cada fila se señala con una “X” las columnas de los procesos cognitivos que el alumnado debe movilizar al responder a la pregunta correspondiente.

Se revisan las pruebas estandarizadas de Matemáticas realizadas en Galicia y Portugal durante los años 2024 y 2025 (12 en Galicia y 16 en Portugal). En Galicia, se analizan las dos evaluaciones de diagnóstico de 9 y 13 años (Xunta de Galicia, 2024; 2025) y las cuatro pruebas de 17 años de Matemáticas y Matemáticas Aplicadas a las Ciencias Sociales (Comisión Interuniversitaria de Galicia, 2026), en adelante *17M* y *17S*. En Portugal, se examinan las cuatro pruebas finales de: 14 años (Instituto de Avaliação Educativa, I. P., 2026b), 16 años – Matemáticas y Matemáticas Aplicadas a las Ciencias Sociales –, y 17 años (Instituto de Avaliação Educativa, I. P., 2026a), en adelante *16M*, *16S* y *17P*. Todas las pruebas tienen dos convocatorias, salvo las evaluaciones de diagnóstico de 9 y 13 años, lo que hace un total de 14 pruebas (6 de Galicia y 8 de Portugal).

Los autores determinan de forma independiente los procesos RBT implicados en cada pregunta y alcanzan un alto nivel de concordancia (κ de Cohen = .844). En los pocos casos de desacuerdo, resuelven por consenso.

Resultados

La Tabla 3 muestra las frecuencias de los procesos cognitivos RBT en las pruebas y las correspondientes al reparto Lange.

Tabla 3
Frecuencias de los procesos RBT y las del reparto Lange (entre paréntesis)

	Años	Recordar	Comprender	Aplicar	Analizar	Evaluar	Crear
Galicia	9	7 (12.2)	19 (10)	21 (7.8)	12 (5.6)	5 (3.3)	0 (1.1)
	13	5 (15.9)	8 (13)	18 (10.1)	15 (7.2)	6 (4.3)	0 (1.5)
	17M*	17 (31.8)	10 (26)	46 (20.2)	23 (14.5)	7 (8.6)	1 (2.9)
	17S*	8 (38.2)	12 (31.2)	57 (24.3)	31 (17.4)	11 (10.4)	6 (3.5)
Portugal	14*	18 (38.9)	18 (31.7)	51 (24.6)	25 (17.7)	13 (10.5)	2 (3.6)
	16M*	5 (31.5)	26 (25.7)	42 (20)	14 (14.3)	16 (8.6)	0 (2.9)
	16S*	1 (31.2)	24 (25.5)	43 (19.8)	25 (14.2)	5 (8.5)	4 (2.8)
	17P*	18 (43.1)	15 (35.6)	54 (27.4)	37 (19.6)	13 (11.7)	4 (3.9)

Fuente: Elaboración propia.

Nota: se señala con * las que poseen doble convocatoria.

Todas las frecuencias de *recordar* son menores que las de Lange, destacando 16M y 16S; en cambio, todas las frecuencias de *aplicar* están entre 1.8 y 2.7 veces las de Lange. Asimismo, no hay diferencia significativa entre las frecuencias de *evaluar* y las correspondientes de Lange ($p = .827$). La distribución de la prueba de 9 años es la única que no difiere significativamente de la de Lange ($p = .110$).

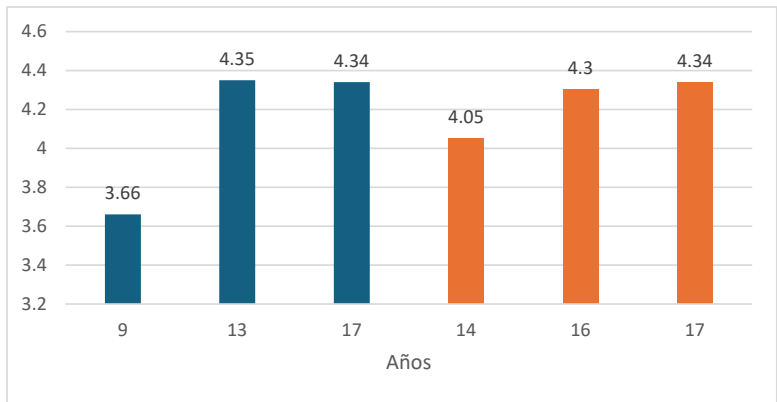
En Galicia, solo hay diferencia significativa entre 9 años y 17M ($p = .031$) y con 17S ($p = .004$). En Portugal, hay diferencia significativa entre todos los años, salvo entre 14 y 17P ($p = .758$). La comparación entre Galicia y Portugal no avala diferencias entre las pruebas de los 17 años ni entre las de 13 y 14, pero sí entre 16S y 17S ($p = .029$).

El Gráfico 1 muestra en azul las demandas cognitivas de las pruebas gallegas y en naranja, las portuguesas.

No existe diferencia estadísticamente significativa entre los valores de *D* ($p = .099$ para 9 y 13 años). La demanda cognitiva aumenta ligeramente con la edad, pero luego se estabiliza. Las demandas se sitúan alrededor de 4 (valor que corresponde a *aplicar*) y ninguna alcanza la media de la escala (5). En Galicia, los valores de *D* para las dos pruebas de los 17 años son bastante dispares (3.92 y 4.70); en Portugal, la diferencia entre las dos pruebas de los 16 años es menor (4.19 y 4.41). La demanda cognitiva media de las pruebas gallegas es 4.16 para un promedio de edad de 14 años y la de las portuguesas, 4.25 para una edad media de 15.75 años.

Para comparar las distribuciones en las preguntas liberadas de PISA 2022 (15 años) con las de las pruebas estandarizadas, dado que las edades no coinciden, se promedian las frecuencias de los 13 y 17 años en Galicia y las de los 14 y 16 en Portugal. Asimismo, se hacen corresponder los dominios cognitivos PISA 2022 con los

Gráfico 1
Valores de D de las pruebas estandarizadas



Fuente: Elaboración propia.

procesos RBT: *formulación con recordar + comprender; empleo con aplicar; interpretación y evaluación con analizar + evaluar y razonamiento con crear*. En ambos casos, se producen desajustes significativos ($p < .001$).

Para comparar las distribuciones de 9, 13 y 14 años con las de las pruebas TIMSS 2023 (la de 14 años con la de 13), se alinean sus dominios cognitivos con los procesos RBT: *conocimiento con recordar + comprender; aplicación con aplicar + analizar y razonamiento con evaluar + crear*. En todos los casos se obtiene diferencia significativa ($p = .033$, $p = .003$ y $p < .001$, respectivamente).

Finalmente, si se asimilaran los dominios cognitivos PISA 2022 y TIMSS 2023: *formulación con conocimiento; empleo, interpretación y evaluación con aplicación; y razonamiento con razonamiento*, promediándose los porcentajes TIMSS 2023 de los 9 y 13 años, resulta que no hay diferencia significativa entre las distribuciones PISA y TIMSS ($p = .130$).

Discusión y conclusiones

Primeramente, respondemos a la pregunta sobre si la evaluación estandarizada se adecúa al aprendizaje a los 9, 13 y 14 años. La demanda cognitiva RBT de los criterios de evaluación en el currículo gallego de 9 años se sitúa en 3.84 (Tugores *et al.*, 2025), solo 0.18 puntos más que la prueba diagnóstica, y la de los criterios de 13 años en 4.88 (Tugores, 2024), 0.53 puntos más que la correspondiente prueba. En Portugal,

la demanda cognitiva de los aprendizajes a los 11 años, calculada a partir de la distribución RBT obtenida por Leite y Valente (2024), es de 5.38, por lo que si en la prueba final que se realiza a los 14 años se requiere 4.05, estamos ante una diferencia mayor.

Esta desalineación está en consonancia con el nivel medio-bajo que Cárdenas *et al.* (2016) identificaron, tras un amplio estudio experimental, en las evaluaciones de Educación Secundaria. Con todo, creemos que estos márgenes no necesitan ser corregidos para que las pruebas estandarizadas gocen de asequibilidad y no provoquen rechazos en los estamentos educativos implicados.

Todas las pruebas presentan una demanda cognitiva mayor que la del modelo Lange. Si este hecho estuviese generalizado, sería una razón para revisar la validez del modelo o, en su defecto, tomar su demanda cognitiva como una cota inferior de exigencia y no como un apropiado valor de equilibrio.

Las pruebas de 14 y 16 años son las que están más próximas, en edad, a las de PISA 2022 (15 años) y sus demandas cognitivas se hallan 1.45 y 1.20 puntos, respectivamente, por debajo de la calculada para PISA 2022, de manera que el alumnado portugués se enfrentó en PISA 2022 a un nivel cognitivo sensiblemente mayor que el exigido en las pruebas nacionales. Sin embargo, todas las pruebas gallegas tienen una demanda cognitiva menor que PISA 2022, aunque mayor que la de PISA en Ciencias a comienzos de milenio ($D = 1.78$) calculada a partir de Rosales *et al.* (2020).

Por otro lado, las demandas cognitivas de las pruebas de 9, 13 y 14 años también son menores que las correspondientes de las pruebas TIMSS 2023 (la de 14 comparada con la de 13). En consecuencia, cabe preguntarse si habría que tener en cuenta estas circunstancias a la hora de enjuiciar las calificaciones obtenidas por el alumnado en las pruebas de carácter internacional.

En España se aplican pruebas de diagnóstico de la competencia STEM (*Science, Technology, Engineering and Mathematics*) a los 11 y 15 años. Los porcentajes de los procesos RBT que deben respetar estas pruebas están recogidos en el documento MGESE 2026. Las demandas cognitivas de las pruebas son: $D = 4.2$ a los 11 años y $D = 5$ a los 15. Ambas son mayores que $DL(6)$ y que las de las pruebas de diagnóstico de 9 y 13 años, aunque los estudiantes son, en general, dos años mayores, y tienen un perfil muy específico.

La demanda cognitiva en la prueba de diagnóstico de 13 años es un 43% mayor que la que resulta del estudio de Caraballo *et al.* (2011) para el curso 2008-09 ($D = 3.05$, apenas por encima de $DL(3)$). Este aumento en quince años evidencia un incremento de la demanda.

En Portugal, que no se aumente la demanda cognitiva en las pruebas del curso previo a la Educación Universitaria podría justificarse por la tensión que generan en muchos estudiantes, pues de sus resultados depende cursar los estudios superiores deseados.

Tanto en Portugal como en Galicia se prioriza una mayor demanda cognitiva en Matemáticas Aplicadas a las Ciencias Sociales que en Matemáticas. Probablemente se deba a que incorporan modelización, toma de decisiones, estadística y probabilidad, e interpretación de fenómenos sociales, lo que exige analizar, evaluar y justificar.

En relación con la distribución de los procesos cognitivos en las pruebas, llama la atención la disparidad del proceso *aplicar* respecto al modelo Lange. Si esta predilección por *aplicar* se entendiera como un contrapeso beneficioso para el equilibrio cognitivo y se trasladara al ámbito teórico, deformaría el triángulo Lange, lo que invitaría a pensar en sustituirlo por otra figura geométrica.

Esta relevancia de *aplicar*, constatada también por Azimi *et al.* (2025), y las dificultades del alumnado para ponerlo en práctica en pruebas estandarizadas descritas por Fernandes (2007), exige proponer un incremento en el aula de tareas que lo desarrollen. Por otro lado, la bajísima frecuencia de *recordar* en las pruebas de 16 años abre el debate sobre la conveniencia de mantener la hoja de fórmulas incluida en las pruebas portuguesas.

Que la distribución de los procesos en las pruebas gallegas sea semejante a la de sus homólogas portuguesas sugiere que el reparto cognitivo estaría ligado a la edad del alumnado, como parece que debe ser. En ambos casos, las distribuciones de PISA 2022 y TIMSS 2023 difieren significativamente de las de las pruebas estandarizadas, por lo que podría resultar apropiado introducir algunos ajustes en estas últimas para aproximar sus perfiles cognitivos.

Los resultados obtenidos en el estudio dialogan con otras áreas de la educación. En el ámbito de la medición educativa, aportan elementos para reflexionar sobre el diseño de instrumentos de evaluación a gran escala. Desde la función docente, invitan a diseñar estrategias para aminorar el impacto cognitivo que pueden generar evaluaciones que priorizan el razonamiento matemático en situaciones cotidianas inéditas. Además, en los debates actuales sobre cómo diseñar una evaluación formativa de calidad, la información derivada de las pruebas estandarizadas proporciona al profesorado un referente para revisar la demanda cognitiva de sus propias evaluaciones.

Por otro lado, las implicaciones pedagógicas de los perfiles obtenidos apuntan también a no relegar las tareas de mayor complejidad. Así, la práctica en el aula

debería articular secuencias en las que lo algorítmico se complemente con lo creativo, de modo que ningún proceso cognitivo quede infrarrepresentado en el día a día.

En la controversia abierta por Boal *et al.* (2008) y López *et al.* (2020), los resultados nos posicionan al lado de los primeros: en los cursos preuniversitarios la incidencia de los procesos cognitivos de nivel superior es baja y *crear* aparece de forma testimonial. Como muestra, ninguna de las pruebas exigió al alumnado generar ideas o proponer estrategias para resolver problemas, algo muy apropiado para valorar su inventiva/creatividad matemática.

El carácter censal de las evaluaciones estandarizadas impide cualquier tipo de diferenciación que contemple factores tangenciales, como los apuntados por Da Silva *et al.* (2020), a la hora de fijar las exigencias cognitivas de las pruebas. Sin embargo, en Matemáticas cabría explorar la incorporación de bloques complementarios con variantes controladas de enunciado que permitan tener en cuenta la diversidad del alumnado.

Por lo que respecta a la formación docente, el patrón de demanda cognitiva observado subraya la importancia de una alfabetización evaluativa: ya no sólo en el diseño de los instrumentos de evaluación, sino también en saber identificar e interpretar evidencias del razonamiento del alumnado y utilizarlas para ofrecer retroalimentación y tomar decisiones didácticas que mantengan el rigor cognitivo de la actividad matemática.

En clave de política educativa, podría ser útil complementar los informes sobre pruebas estandarizadas con indicadores de demanda cognitiva del instrumento. Esta metainformación haría posible contextualizar comparaciones y evitar lecturas lineales de las puntuaciones del alumnado.

Para una audiencia internacional, nuestra principal aportación es metodológica: el procedimiento de codificación mediante RBT y el cálculo de la demanda cognitiva permiten replicar el análisis y detectar desajustes entre currículo, evaluación nacional y marcos internacionales. Asimismo, facilitan considerar diferentes finalidades evaluativas desde una perspectiva cognitiva, como ponderar unos procesos frente a otros.

La transferencia de los hallazgos a la planificación de los centros educativos puede concretarse incorporando la promoción de determinados procesos cognitivos como criterio adicional para asignar recursos, organizar el acompañamiento pedagógico y seleccionar materiales didácticos.

El estudio realizado presenta ciertas limitaciones metodológicas relacionadas con el corpus analizado por el número de pruebas o el carácter geográfico de las mismas, lo que dificulta la generalización de los resultados. Además, no se ha examinado la

relación entre la demanda cognitiva y variables como la dificultad matemática o las puntuaciones obtenidas por el alumnado.

Para minimizar posibles sesgos en la clasificación, se consideraron aspectos como el número de pasos de procesamiento, el nivel de abstracción y las conexiones entre conceptos. Sin embargo, la codificación no contempla las estrategias de resolución que haya podido adoptar el alumnado.

Al mismo tiempo, el diseño de las pruebas estandarizadas es altamente heterogéneo. Por ejemplo, Galicia y Portugal no emplean la demanda cognitiva como criterio de balance, a diferencia de lo que ocurre en Estados Unidos (*National Assessment Governing Board*, 2026).

Como línea prospectiva, la ampliación de este estudio a otros sistemas educativos permitiría elaborar repartos de demanda cognitiva ajustados a las distintas realidades nacionales. Además, podría proponerse un triángulo cognitivo acorde con las prácticas actuales, con regiones de tamaño distinto al asignado en el reparto Lange. Asimismo, podrían elaborarse mapas nacionales de demanda cognitiva por edades del alumnado. Aunque la recogida de información a lo largo de más años escolares debería contemplar eventuales modificaciones del currículo, adaptaciones de las pruebas derivadas del rendimiento del alumnado y, en última instancia, decisiones extraacadémicas.

Referencias bibliográficas

- Anderson, L. W., & Krathwohl, D. R. (2001). *A taxonomy for learning, teaching and assessing: A revision of Bloom's taxonomy of educational objectives*. Addison Wesley Longman.
- Auld, E., & Morris, P. (2013). Comparative education, the 'new paradigm' and policy borrowing: Constructing knowledge for educational reform. *Comparative Education*, 50(2), 129-155. <https://doi.org/10.1080/03050068.2013.826497>
- Azimi, S., Gunpinar, Y., Atabas, S., & Zolfaghari, M. (2025). Strands and cognitive demand levels: Examining university entrance exam questions across three countries. *Journal of Mathematics and Science Teacher*, 5(3), em084. <https://doi.org/10.29333/mathsciteacher/16847>
- Boal, N., Bueno, C., Lerís, M. D., & Sein-Echaluce, M. L. (2008). Las habilidades matemáticas evaluadas en las pruebas de acceso a la universidad: un estudio en varias universidades públicas españolas. *Revista de Investigación Educativa*, 26(1), 11-23. <https://revistas.um.es/rie/article/view/94081/90711>
- Boertien, H., & De Lange, J. (1994). *The national options of TIMSS in The Netherlands*. University of Twente.
- Caraballo, R. M., Rico, L., & Lupiáñez, J. L. (2011). Análisis de los ítems de las evaluaciones autonómicas de diagnóstico en España: 2008-2009. *Unión: Revista Iberoamericana de*

- Educación Matemática*, 7(26), 27-40. <https://www.revistaunion.org/index.php/UNION/article/view/925/629>
- Cárdenas, J. A., Lorenzo, B. N., & Cáceres, M. J. (2016). La evaluación de las matemáticas: análisis de las pruebas escritas que se realizan en la secundaria. *Unión: Revista Iberoamericana de Educación Matemática*, 12(48), 59-78. <https://www.revistaunion.org/index.php/UNION/article/view/527/280>
- Comisión Interuniversitaria de Galicia. (2026). *Proba de avaliación de bacharelato para o acceso á universidade*. <https://ciug.gal/pau>
- Da Silva, D., Vaz-Rebelo, P., & Canhoto, C. (2020). Avaliação adequada ao currículo? O que dizem os conteúdos solicitados nas provas de Biologia dos exames nacionais em Portugal e no Brasil. *Ensaio: Pesquisa em Educação em Ciências*, 22, 1-20. <https://dx.doi.org/10.1590/1983-21172020210106>
- Fernandes, D. (2007). A avaliação das aprendizagens no sistema educativo português. *Educação e Pesquisa*, 33(3), 581-600. <https://doi.org/10.1590/S1517-97022007000300013>
- Fernández, M., Alcaraz, N., & Sola, M. (2017). Evaluación y pruebas estandarizadas: una reflexión sobre el sentido, utilidad y efectos de estas pruebas en el campo educativo. *Revista Iberoamericana de Evaluación Educativa*, 10(1), 51-67. <https://doi.org/10.15366/riee2017.10.1.003>
- Gutiérrez, J. G., & Acuña, L. A. (2022). Evaluación estandarizada de los aprendizajes: una revisión sistemática de la literatura. *CPU-e. Revista de Investigación Educativa*, 34, 321-351. <https://dialnet.unirioja.es/descarga/articulo/8349965.pdf>
- Hernández-Sampieri, R., Fernández-Collado, C., & Baptista-Lucio, P. (2014). *Metodología de la investigación* (6.ª ed.). McGraw-Hill.
- Instituto de Avaliação Educativa, I. P. (2022). *Estudo de aferição amostral do Ensino Básico*. <https://bit.ly/4sj2Hxg>
- Instituto de Avaliação Educativa, I. P. (2026a). *Provas e exames finais nacionais ES*. <https://bit.ly/4aSi3mw>
- Instituto de Avaliação Educativa, I. P. (2026b). *Provas finais de ciclo EB*. <https://bit.ly/4b9UXXw>
- Instituto Nacional de Evaluación Educativa. (2022). *Marcos de evaluación TIMSS 2023*. Ministerio de Educación y Formación Profesional. <https://bit.ly/3N4ds7G>
- Leite, T., & Valente, B. (2024). A exigência cognitiva do currículo dos primeiros anos de escolaridade em Portugal: uma análise comparativa. *Revista Lusófona de Educação*, 61, 29-47. <https://revistas.ulusofona.pt/index.php/rleducacao/article/view/9317>
- López-Beltrán, M., Albarracín, L., Ferrando, I., Montejo-Gámez, J., Ramos, P., Serradó, A., Thibaut, E., & Mallavibarrena, R. (2020). La educación matemática en las enseñanzas obligatorias y el Bachillerato. En D. Martín de Diego (Coord.), *Libro Blanco de las Matemáticas* (pp. 1-94). Centro de Estudios Ramón Areces. <https://bit.ly/4dMiVYC>
- Marques, M. M., Sousa, J., Costa, N., & Pacheco, J. A. (2015). Efeitos da avaliação externa das aprendizagens no desenvolvimento profissional de professores de Matemática do Ensino

- Básico em Portugal. *Meta: Avaliação*, 7(19), 58-84. <https://repositorium.sdum.uminho.pt/handle/1822/36483>
- Martínez, F. (2009). Evaluación formativa en aula y evaluación a gran escala: hacia un sistema más equilibrado. *Revista Electrónica de Investigación Educativa*, 11(2), 1-18. <http://redie.uabc.mx/vol11no2/contenido-mtzrizo2.html>
- Ministerio de Educación, Formación Profesional y Deportes. (2023). *PISA 2022: Programa para la Evaluación Internacional de los Estudiantes. Informe Español*. <https://bit.ly/3Cos641>
- Ministerio de Educación, Formación Profesional y Deportes. (2026). *Marco general de las evaluaciones del sistema educativo. Evaluación general del sistema y evaluaciones de diagnóstico* (2.ª ed.). <https://bit.ly/47sy6VO>
- National Assessment Governing Board. (2026). *2026 NAEP Mathematics Framework*. U.S. Department of Education. <https://www.nagb.gov/naep-subject-areas/mathematics/2026-naep-mathematics-framework.html>
- Pizà-Mir, B. (2022). Validation of the use of Bloom's revised taxonomy as a tool for the design of assessment tests. En I. Cruz, J. L. Mateu, & C. Méndez (Coords.), *Innovación educativa para el desarrollo sostenible, la economía y la empresa* (pp. 904-925). Aula Magna McGraw Hill. <https://doi.org/10.20944/preprints202212.0123.v2>
- Popham, W. J. (1999). Why standardized test scores don't measure educational quality. *Educational Leadership*, 56(6), 2-11. https://www.researchgate.net/publication/242489679_Why_Standardized_Tests_Don't_Measure_Educational_Quality
- Rodríguez, P., Soca, J., Castillo, M., & Luzardo, M. (2025). Evaluaciones educativas estandarizadas: desde el marco a los resultados. *Cuadernos de Investigación Educativa*, 16(2). <https://doi.org/10.18861/cied.2025.16.2.4015>
- Rosales, E. M., Rodríguez, P. G., & Romero, M. (2020). Conocimiento, demanda cognitiva y contextos en la evaluación de la alfabetización científica en PISA. *Revista Eureka sobre Enseñanza y Divulgación de las Ciencias*, 17(2), 2302. https://dx.doi.org/10.25267/Rev_Eureka_ensen_divulg_cienc.2020.v17.i2.2302
- Sarmiento, J. A., & Sarmiento, E. (2023). El currículo de primaria desde la LOMLOE: análisis de las relaciones entre sus elementos. *Investigación: Cultura, Ciencia y Tecnología*, 29, 61-66. <https://dialnet.unirioja.es/servlet/articulo?codigo=9174001>
- Shepard, L. A. (2006). Classroom assessment. En R. L. Brennan (Ed.), *Educational measurement* (4.ª ed., pp. 623-646). Praeger.
- Tugores, L. (2024). Niveles de exigencia cognitiva en tareas de matemáticas (ESO). [Tesis Doctoral]. Universidade de Vigo. <http://hdl.handle.net/11093/6819>
- Tugores, L., Aires, A. P. y Raposo-Rivas, M. (2025). La demanda cognitiva en los currículos de Matemáticas de 6 a 12 años en España y Portugal: un análisis comparativo. *Revista Lusófona de Educação*, 65(65), 131-147. <https://doi.org/10.24140/issn.1645-7250.rle65.06>
- Xunta de Galicia. (2024). *Avaliación de diagnóstico. Competencia matemática. Curso 2023-2024. 4.º curso de Educación Primaria y 2.º curso de Educación Secundaria Obligatoria*.

Xunta de Galicia. (2025). *Avaliación de diagnóstico. Competencia matemática. Curso 2024-2025. 4.º curso de Educación Primaria y 2.º curso de Educación Secundaria Obligatoria.*

Laia Tugores

Universidade de Vigo, Facultad de Educación y Trabajo Social,
Departamento de Matemáticas, España
laia.francina.tugores@uvigo.gal
<https://orcid.org/0000-0001-7679-7683>

Ana Paula Aires

Universidade de Trás-os-Montes e Alto Douro, Portugal
CIDTFF – Centro de Investigação Didáctica e Tecnologia na Formação de
Formadores, Universidade de Aveiro, Portugal
aares@utad.pt
<https://orcid.org/0000-0001-8138-3776>

Manuela Raposo Rivas

Universidade de Vigo, Facultad de Educación y Trabajo Social,
Departamento de Didáctica, Organización Escolar y Métodos de
Investigación, España
mraposo@uvigo.gal
<https://orcid.org/0000-0001-7781-7818>

Data de submissão: Fevereiro, 2026

Data de avaliação: Março-Abril, 2026

Data de publicação: Junho, 2026