Validación de una prueba para diagnosticar destrezas de pensamiento en estudiantes de quinto grado

- Validation of a Test to Diagnose Thinking Skills in Fifth-Grade Students
- Validação de um teste para diagnosticar habilidades de pensamento em estudantes do quinto ano

Maria-Antonia Manassero-Mas*
Ángel Vázquez-Alonso**

Forma de citar este artículo

Manassero-Mas, M. A. y Vázquez-Alonso, Á. (2026). Validación de una prueba para diagnosticar destrezas de pensamiento en estudiantes de quinto grado. *Tecné, Episteme y Didaxis, TED*, (59), 305 - 323. https://doi.org/10.17227/ted.num59-22994

* Doctora en Psicología. Catedrática de Universidad de Psicología Social, Departamento de Psicología, Universidad de las Islas Baleares, Palma, España. ma.manassero@uib.es

** Doctor en Filosofía y Ciencias de la Educación. Investigador honorífico, Instituto de Investigación e Innovación Educativa, Universidad de las Islas Baleares, Palma, España. angel.vazquez@uib.es

Resumen

El pensamiento crítico (PC) es una competencia transversal en la educación del siglo XXI, especialmente importante para la educación científica, matemática y tecnológica, aunque su investigación es desigual, apareciendo más pobre en su evaluación y en educación primaria. Para compensar estas carencias, el objetivo de este estudio es diseñar un test de evaluación de PC dirigido a estudiantes de quinto grado de primaria y presentar sus propiedades psicométricas. La metodología se ajusta a las prescripciones del desarrollo de test, elaborando una prueba de 18 ítems y cuatro destrezas de PC independientes de conocimientos previos, aplicándola a una muestra de 146 estudiantes de quinto grado de primaria y analizando la bondad del ajuste de diversas soluciones factoriales con las respuestas de los estudiantes. La prueba muestra un buen índice de dificultad intermedio y el análisis factorial confirmatorio valida una solución de cuatro factores y 16 ítems que alcanza el mejor ajuste, permite interpretar razonablemente los cuatro factores empíricos frente a las destrezas de PC postuladas teóricamente y presenta los excelentes parámetros psicométricos de la bondad de ajuste que respaldan la confiabilidad y validez de la prueba. Finalmente, se discuten las propiedades y utilidad de la prueba, las limitaciones derivadas del escaso número de ítems en algunos factores, junto a las posibles mejoras futuras para superarlas, así como las aplicaciones de la prueba para evaluar las destrezas de PC en educación e investigación.

Palabras clave

test de pensamiento crítico; validez; fiabilidad; educación primaria

Artículo de investigación

Fecha de recepción: 25/03/2025
Fecha de aprobación: 27/09/2025
Fecha de publicación: 01/01/2026

Primer semestre de 2026 / ISSN 2665-3184
pp. 305 - 323

Abstract

Critical thinking (CT) is a transversal competence in 21st-century education, particularly relevant to science, mathematics, and technology education. However, research in this area remains uneven, with limited progress in assessment—especially at the primary level. To address this gap, the aim of this study was to design a CT assessment test for fifth-grade students and to present its psychometric properties. Following the standard test-development procedures, an 18-item test was created to measure four CT skills independent of prior knowledge. The test was administered to a sample of 146 fifth-grade students, and the goodness of fit of several factorial solutions was examined using their responses. The test demonstrated an appropriate intermediate level of difficulty, and confirmatory factor analysis supported a four-factor solution with 16 items, which achieved the best fit. The four empirical factors were found to align well with the theoretically proposed CT skills, and the psychometric indicators confirmed strong reliability and validity. Finally, the paper discusses the test's properties and usefulness, limitations related to the small number of items in some factors, possible future improvements, and potential applications for assessing CT skills in educational and research contexts.

Keywords

critical thinking test; validity; reliability; primary education

Resumo

O pensamento crítico (PC) é uma competência transversal na educação do século XXI, especialmente relevante para o ensino de ciências, matemática e tecnologia. No entanto, a pesquisa nessa área é desigual, apresentando lacunas importantes na avaliação — particularmente no ensino fundamental. Com o objetivo de reduzir essas lacunas, este estudo teve como propósito desenvolver um teste de avaliação de PC voltado a estudantes do quinto ano do ensino fundamental e apresentar suas propriedades psicométricas. Seguindo as diretrizes para o desenvolvimento de testes, foi elaborada uma prova com 18 itens que avaliam quatro habilidades de PC independentes de conhecimentos prévios. O instrumento foi aplicado a uma amostra de 146 estudantes do quinto ano, e analisou-se o ajuste de diferentes soluções fatoriais com base nas respostas dos participantes. O teste apresentou um nível de dificuldade intermediário adequado, e a análise fatorial confirmatória validou uma solução de quatro fatores e 16 itens, que obteve o melhor ajuste. Os quatro fatores empíricos puderam ser interpretados de forma coerente com as habilidades teóricas de PC postuladas, apresentando excelentes índices psicométricos de ajuste, o que respalda a confiabilidade e a validade do teste. Por fim, discutem-se as propriedades e a utilidade do instrumento, as limitações decorrentes do número reduzido de itens em alguns fatores, as possibilidades de aprimoramento futuro e as aplicações do teste para avaliar habilidades de PC em contextos educacionais e de pesquisa.

Palavras-chave

teste de pensamento crítico; validade; confiabilidade; ensino fundamental

Introducción

Diversas organizaciones respaldan las destrezas del siglo XXI porque consideran que esta compleja competencia es necesaria para afrontar los desafíos de las actuales sociedades del conocimiento (European Union, 2014; International Society for Technology Education, 2003; National Research Council, 2012; Organisation for Economic Co-operation and Development [OECD], 2018; Unesco, 2016).

Aunque la definición de las destrezas varía según la fuente, suelen comprender destrezas digitales y cognitivas. Estas últimas, a su vez, se distinguen entre destrezas blandas (psicosociales o interpersonales) y duras (destrezas cognitivas de orden superior). Algunos autores (Fullan y Scott, 2014) las denominan también modelos de 4C (colaboración, comunicación, creatividad y pensamiento crítico) o 6C (añaden carácter y ciudadanía). El hecho central es que el pensamiento crítico (PC) aparece como componente invariable de las destrezas del siglo XXI, lo cual proyecta demandas directas sobre su educación (Almerich *et al.*, 2020; Vincent-Lancrin *et al.*, 2019).

Por otro lado, las encuestas laborales muestran que trabajadores y empresarios valoran el PC como el principal requerimiento de los futuros trabajos (World Economic Forum, 2021) y como un factor clave para el éxito de las personas en la era de la ciencia y la información (Tremblay, 2013). Además, el lento desarrollo cognitivo humano, junto con la necesidad laboral del PC, ha condicionado que la mayoría de los esfuerzos por educar el PC se hayan concentrado en la educación superior, en detrimento de los niveles educativos anteriores, para los cuales el PC constituye aún una innovación y un reto pendientes (Aktoprak y Hursen, 2022).

Los estudios pioneros de Piaget (Piaget e Inhelder, 1997) y los programas de aceleración cognitiva (Shayer y Adey, 2002) han demostrado empíricamente el impacto relevante y positivo de las destrezas cognitivas en el aprendizaje. La revisión de Hattie (2009) ha confirmado que el tamaño del efecto de los programas piagetianos en el aprendizaje es muy grande, y el impacto de diferentes destrezas específicas de PC también es considerable, de modo que el dominio de las destrezas de PC se considera un factor clave para lograr el aprendizaje significativo y profundo que caracteriza la excelencia educativa (Hattie, 2012; O'Hare y McGuinness, 2015; Valenzuela, 2008).

En suma, el pensamiento crítico tiene un impacto beneficioso en el aprendizaje, el éxito profesional y laboral, y la ciudadanía democrática del siglo XXI, lo cual justifica la atención innovadora hacia el PC como elemento central de la educación en todos los niveles escolares.

Sin embargo, la mayor parte de la investigación sobre PC se ha centrado en planear métodos de enseñanza efectivos, cuyo desarrollo, principalmente cualitativo, ha descuidado la evaluación cuantitativa de los logros (Gelerstein *et al.*, 2016). Por niveles, la mayoría de estudios se han desarrollado en la universidad, hay pocos en secundaria y apenas existen en primaria (Aktoprak y Hursen, 2022). Para paliar la escasez de evaluación cuantitativa del PC en la educación primaria, este estudio aporta la construcción y validación psicométrica de un test para evaluar cuatro destrezas de PC en estudiantes de primaria (10-11 años).

Antecedentes

La nota más conocida de la investigación sobre el PC es la ausencia de consenso en algo tan básico como una definición compartida entre los especialistas, debido a la diversidad de enfoques filosóficos (Ennis, 2018; Facione, 1990; Colom *et al.*, 2014) y psicológicos

(Bailin *et al.*, 1999; Halpern, 2003; Lai, 2011). Hace años, Ennis (2018) definió el PC como un pensamiento reflexivo y razonable centrado en decidir qué creer o hacer; posteriormente desarrolló las disposiciones y destrezas involucradas en tales decisiones, y es uno de los autores más citados. Un panel de la American Philosophical Association acordó una definición de PC como el juicio deliberado y autorregulado con un propósito específico, que emplea interpretación, análisis, evaluación e inferencia basados en evidencia, conceptos, métodos, criterios y contextos para establecer dichos juicios, definición que es una referencia obligada por su elaboración consensuada (Facione, 1990).

Desde la psicología cognitiva se conceptualiza el PC subrayando su composición de múltiples habilidades cognitivas de orden superior (razonar, decidir, resolver, evaluar, etc.), disposiciones actitudinales (curiosidad, apertura de mente, sistematicidad, búsqueda de la verdad, etc.) y estándares de calidad (precisión, solidez, coherencia, relevancia, adecuación, etc.), como herramientas para superar las tendencias naturales del pensamiento al error, la falacia y el sesgo (egocentrismo y sociocentrismo) y producir juicios válidos. Estas destrezas, disposiciones, normas y valores inherentes al PC aportan una base crucial para su evaluación (Bailin *et al.*, 1999).

La evidencia de que el constructo PC está compuesto por múltiples destrezas específicas, mencionadas en los párrafos anteriores, constituye quizá el acuerdo más generalizado sobre el PC. Por ello, como alternativa a la falta de consenso conceptual, algunos especialistas —en especial los constructores de test— definen el PC por extensión, es decir, especificando sus destrezas constitutivas (Fisher, 2009). La citada definición de la APA (Facione, 1990) asume también este enfoque extensivo al mencionar algunas destrezas (interpretación, análisis, evaluación, inferencia, juicio y autorregulación), y el desarrollo ampliado de Ennis (2018) incluye otras muchas, entre las que destacan la resolución de problemas y la toma de decisiones.

Este enfoque extensivo es especialmente evidente en los instrumentos de evaluación del PC desarrollados por los expertos, pues, por su propia naturaleza funcional y práctica, cada instrumento especifica las destrezas que valora como núcleo de su operatividad. Por ello, los tests resultan más concretos que las diferentes definiciones. Por ejemplo, el *Cornell Critical Thinking Tests Level X* (Ennis y Millman, 2005a) valora cinco destrezas (inducción, deducción, observación, credibilidad y supuestos), y el test *Critical Thinking Assessment* de Halpern (2010) valora análisis argumental, comprobación de hipótesis, probabilidad e incertidumbre, resolución de problemas y razonamiento verbal. Sin embargo, la comparación entre diversos instrumentos de evaluación muestra también una falta de coincidencia en las destrezas evaluadas por unos y otros, aunque algunas (análisis, razonamiento, resolución de problemas, toma de decisiones) y disposiciones (mente abierta, curiosidad) son predominantes (Manassero-Mas *et al.*, 2022; Ennis y Chattin, 2018; Lai, 2011).

En resumen, la investigación sobre el PC muestra una ausencia de consenso acerca de su conceptualización, consecuencia de la complejidad inherente al constructo y no tanto un defecto de la investigación. Sin embargo, esta falta de consenso se transmite también a su enseñanza y evaluación (Saiz, 2017).

Para afrontar dicha complejidad, se han propuesto en los últimos años algunas taxonomías globales que intentan sistematizar el campo. Así, Dwyer *et al.* (2014) desarrollaron un marco que integra objetivos educativos, procesos cognitivos (juicio reflexivo, autorregulación y metacognición) y destrezas de PC (análisis, evaluación e inferencia), junto con los necesarios procesos de memoria y comprensión para su aplicación. Posteriormente se han elaborado dos taxonomías más sencillas, que muestran varias coincidencias entre sí, como organizar el PC en cuatro dimensiones. Vázquez-Alonso y Manassero-Mas (2018) propusieron cuatro dimensiones básicas del PC: creatividad, razonamiento y argumentación, procesos complejos y juicio y evaluación; cada una se desarrolla, a su vez, en múltiples categorías y subcategorías (por ejemplo, los procesos complejos comprenden la resolución de problemas y la toma de decisiones). De forma análoga, Fisher (2021) organizó las destrezas de PC en cuatro grupos básicos (interpretación, análisis, evaluación y autorregulación), cuyos contenidos coinciden ampliamente con las dimensiones de la taxonomía anterior.

La taxonomía de Vázquez-Alonso y Manassero-Mas (2019) se toma como referencia en este estudio, donde el constructo PC se considera fundamental y constituye el primer nivel respecto a las cuatro dimensiones mencionadas, cada una de las cuales contiene múltiples destrezas de pensamiento específicas y otros conceptos asociados (disposiciones y normas de actitud), que proporcionan las destrezas de PC involucradas en los instrumentos de evaluación del pensamiento crítico.

Marco teórico: la evaluación del pensamiento

La declaración de expertos de la APA (Facione, 1990) recomendó complementar la enseñanza del PC con su evaluación frecuente y explícita (recomendaciones 12 y 13), tanto diagnóstica como sumativa, utilizando instrumentos válidos, confiables y equitativos, hoy día requisitos obvios en cualquier prueba o test de evaluación (Muñiz y Fonseca-Pedrero, 2019). Ennis (2018) justifica la necesidad de evaluar el PC con múltiples razones que reflejan diferentes fines teóricos de la evaluación educativa: diagnosticar el nivel del alumnado, retroalimentar el progreso, motivar el aprendizaje del PC, informar a los docentes sobre su enseñanza, investigar el PC, asesorar en la elección de estudios y estimular a las instituciones educativas a reportar sus resultados de enseñanza. Estas razones convergen con las funciones generales asignadas a la evaluación educativa, cuyo logro requiere la construcción de instrumentos válidos y confiables para medir el PC (Szőköl *et al.*, 2022).

La mayoría de los instrumentos de evaluación del PC se orienta a medir unas pocas destrezas, cuyo número y naturaleza varían entre los diferentes instrumentos (Facione *et al.*, 1998; Halpern, 2010; Watson y Glaser, 2002), aunque todos guardan correspondencia con las taxonomías de PC mencionadas (Vázquez-Alonso y Manassero-Mas, 2019; Fisher, 2021).

Por otro lado, la gran mayoría de los instrumentos de evaluación existentes se dirige a adultos y estudiantes universitarios, y apenas existen pruebas específicas para estudiantes jóvenes. Algunos test permiten adaptaciones o incluyen partes aplicables a diferentes edades,

como es el caso de los test de Cornell (X, Y, Z) (Ennis y Millman, 2005a, 2005b) y de otras propuestas de investigación que aún carecen de validación (Lopes *et al.*, 2018). En particular, la revisión de Aktoprak y Hursen (2022) reportó la gran escasez de investigaciones sobre PC en educación primaria y evidenció que las existentes muestran un predominio de metodologías cualitativas de evaluación (Gelerstein *et al.*, 2016). Por tanto, se propone una mayor orientación hacia la educación elemental y las metodologías cuantitativas, que ofrezcan mediciones confiables y comparables del PC. En la misma línea, Wang y Chen (2024) también han señalado la escasez de estudios en primaria relativos a la investigación de las disposiciones del PC y han elaborado su propia propuesta.

En suma, el desarrollo de la investigación sobre el PC ha sido muy desigual entre los distintos niveles educativos —amplio en la universidad y más escaso cuanto menor es el nivel— y, en cuanto a sus contenidos, predomina la enseñanza, mientras que la evaluación fiable es escasa (especialmente entre los estudiantes más jóvenes). Estas carencias justifican la atención de este estudio hacia la evaluación del PC en los más jóvenes, centrada en destrezas específicas, funcionales y adaptadas a su desarrollo cognitivo, de modo que contribuya a diagnosticar y visibilizar el PC en niveles tempranos.

El objetivo de este estudio es desarrollar un test diseñado para evaluar el PC de estudiantes de quinto grado y validar sus características psicométricas, de manera que contribuya a subsanar la carencia de pruebas en educación primaria. El propósito principal es identificar psicométricamente los factores subyacentes que representen de forma óptima las respuestas de los estudiantes encuestados y las distintas destrezas del constructo PC, así como validar la estructura y consistencia interna del instrumento; no se abordan, en cambio, otros tipos de validez del instrumento.

Metodología

Las recomendaciones para desarrollar test fiables (Muñiz y Fonseca-Pedrero, 2019) se aplicaron a una serie de estudios piloto previos con bancos de ítems de PC (Manassero-Mas y Vázquez-Alonso, 2020a, 2020b) y a la validación de un test para sexto grado de primaria (Manassero-Mas y Vázquez-Alonso, 2023, 2024). Estas recomendaciones se aplican aquí con la metodología descrita en esta sección para elaborar un test dirigido a estudiantes de quinto grado de primaria y determinar su validez y fiabilidad.

Instrumento

El test *Retos de Pensamiento* (RdP_EP5), validado en este estudio, consta de 18 ítems que evalúan cuatro destrezas de PC. La destreza *clasificación* (dimensión de creatividad) evalúa la capacidad de agrupar o separar varios elementos según la valoración de sus rasgos comunes o diferenciales. El *razonamiento lógico* (dimen-

sión de razonamiento) evalúa la capacidad deductiva simple (silogismo simple). La *toma de decisiones* y la *resolución de problemas* (dimensión de procesos complejos) miden la capacidad de identificar las mejores decisiones o soluciones en una situación particular. Estas destrezas se acordaron con los colegios participantes en el estudio, teniendo en cuenta, además, su adaptación cognitiva a la edad y a los aprendizajes habituales en quinto grado de educación primaria (EP5).

Los investigadores diseñaron los ítems del test RdP_EP5 atendiendo a criterios de sencillez lectora, facilidad de comprensión, planteamiento de un desafío interesante y motivador para los estudiantes, y concordancia entre la demanda cognitiva del ítem y el desarrollo evolutivo del alumnado destinatario. Los ítems plantean una variedad de escenarios y situaciones, y comunican su información mediante representaciones predominantemente figurativas, sobre las cuales se formulan una o varias preguntas. La demanda cognitiva de cada pregunta se ajusta a la destreza que representa, al desarrollo cognitivo de los estudiantes y al propósito de plantear un reto de pensamiento auténtico y estimulante. Los investigadores asignaron cada ítem a la destreza que teóricamente mostraba mayor congruencia con su contenido (Tabla 1).

Tabla 1.

Especificaciones del test aplicado (RdP_EP5) en este estudio para evaluar las destrezas de pensamiento en quinto grado de educación primaria EP5

Destrezas de pensamiento	Fuente	Tipo	Ítems	Fiabilidad (Alfa / Omega)	
Clasificación (CLAS)	Elaboración del autor*	Figurativo	5	0,615	0,621
Resolución de problemas (PROB)	Elaboración del autor*	Figurativo	7	0,663	0,679
Toma de decisiones (DECIS)	Elaboración del autor*	Figurativo	4	0,563	0,597
Razonamiento lógico (LOG-RA)	Ennis y Millman, 2005b	Verbal	2	0,392	0,393
Total			18	0,736	0,740

* Traducidos y adaptados de materiales abiertos de https://www.criticalthinking.com

Fuente: elaboración propia.

El test RdP_EP5 es una prueba libre de cultura, ya que el contenido de los ítems, por diseño, es independiente de los contenidos escolares y la búsqueda de la respuesta correcta no requiere conocimientos previos, sino únicamente razonar sobre la información presentada. De este modo, las respuestas son independientes del conocimiento previo, a diferencia de la mayoría de los test de PC. Por ejemplo, para responder correctamente a la prueba *Science* CT se requieren conocimientos del currículo de ciencias de primaria (Mapeala y Siew, 2015).

Los formatos de respuesta del RdP_EP5 son mayoritariamente cerrados (aunque ocho ítems solicitan una respuesta abierta breve), pues permiten una evaluación estandarizada, rápida, válida y confiable del test, así como establecer líneas diagnósticas de base para comparar diferentes investigaciones, programas y metodologías de enseñanza. La Tabla 1 presenta la estructura y los coeficientes de fiabilidad del test y de sus cuatro destrezas teóricas.

Participantes

Tras la depuración de las respuestas de los estudiantes al test, la muestra válida quedó compuesta por 146 estudiantes de quinto

grado de primaria (88 niños y 58 niñas), con una edad media de 10,3 años, pertenecientes a seis colegios españoles. La mayoría de los centros (61,6 %) son públicos y están ubicados en diferentes tipos de poblaciones (grandes, medianas y pequeñas) y contextos sociales (centro, barriadas, etc.). Los colegios participaron en el estudio por su interés en la educación del PC, por lo que conforman una muestra de conveniencia.

Procedimientos

Los estudiantes respondieron el test RdP_EP5 en sus grupos de clase, como una actividad ordinaria y reglada de evaluación educativa del aprendizaje escolar dirigida por sus docentes, con el propósito de estimular el esfuerzo y la motivación, y sin participación de los investigadores. Los docentes aplicaron el test digitalmente, siguiendo pautas estandarizadas comunes en todos los colegios y sin límite de tiempo dentro de un período de clase, suficiente para completarlo.

Las respuestas correctas recibieron un punto, las incorrectas, cero, y no se aplicaron correcciones por respuestas aleatorias. La puntuación de cada destreza o factor se obtuvo sumando las respuestas correctas de los ítems que la componían, y la suma total de aciertos valoró el nivel global de PC de los estudiantes.

La validez de contenido del test RdP_EP5 se fundamenta en la credibilidad académica de las publicaciones especializadas consultadas (Tabla 1) y en el juicio profesional de los investigadores al seleccionar, elaborar y adaptar los ítems, aplicando el criterio de mejor ajuste entre el contenido del ítem y la destreza asignada, así como entre la demanda cognitiva del ítem y el desarrollo evolutivo de los estudiantes.

Análisis de resultados

Las puntuaciones individuales se procesaron con Jamovi (versión 2.4.11), que evalúa la confiabilidad mediante varios índices (omega o alfa de Cronbach). Las diferencias entre grupos se analizan mediante la significación estadística y el estadístico épsilon cuadrado (ANOVA), que mide la magnitud de la proporción de varianza compartida entre grupos (tamaño del efecto).

La estructura de las respuestas se exploró mediante análisis factorial exploratorio (AFE) con el fin de obtener información básica sobre las relaciones entre variables observadas, identificar sus agrupaciones (con referencia en la estructura teórica del instrumento) y explicar la contribución de las respuestas al constructo latente bajo estudio (destrezas de PC). Para la extracción de factores se utilizó el método de residuos mínimos en combinación con una rotación *oblimin* (que asume factores latentes correlacionados) y el análisis paralelo.

Cuantos más ítems se asocian a un factor, más determinado queda este, su medida es más precisa y la solución factorial más estable; por ello, algunos

autores proponen un mínimo de cuatro variables con cargas sustanciales para identificar claramente un factor (Ferrando y Lorenzo-Seva, 2018). Dada la pequeña longitud del test (para evitar el cansancio en las respuestas), la validación de la estructura busca un equilibrio entre el número de ítems por factor y el número de factores del modelo, procurando evitar tanto estructuras excesivamente complejas como aquellas con factores simples compuestos por ítems teóricamente incoherentes.

El proceso de exploración de modelos implicó la eliminación de los ítems defectuosos que no contribuían significativamente, lo cual exigió analizar el efecto de cada eliminación en la redistribución de las cargas de los ítems restantes y proceder así en las sucesivas estructuras exploradas (Lloret-Segura *et al.*, 2014). Los factores empíricos latentes identificados en cada modelo explican los patrones de correlaciones entre ítems y se utilizan para investigar las relaciones entre las variables latentes y las puntuaciones observadas en las escalas (cargas factoriales, correlaciones, coeficientes estandarizados e índices de modificación), con el objetivo de determinar el grado en que las variables latentes del modelo están representadas por un conjunto particular de ítems.

Los modelos propuestos a partir de las soluciones factoriales se estimaron mediante el método de máxima verosimilitud con varianza de factor ajustada a la unidad. La solución factorial óptima se valoró comparativamente a través de los índices estadísticos de bondad de ajuste general a los datos observados, propios de los modelos de ecuaciones estructurales (SEM, por sus siglas en inglés), en cada modelo empírico o solución factorial. Los índices y valores de corte aplicados (Boateng *et al.*, 2018) fueron los siguientes: 1) el estadístico ji cuadrado (χ^2, $p > 0{,}05$); 2) el cociente ji cuadrado/grados de libertad ($\chi^2/gl \leq 5$); 3) RMSEA ($\leq 0{,}05$) y su intervalo de confianza (IC) del 95 %; 4) SRMR ($\leq 0{,}08$); 5) CFI y TLI (ambos $\geq 0{,}95$); y (6) los valores mínimos de los criterios de información de Akaike (AIC) y bayesiano (BIC).

Resultados y análisis

Estadística descriptiva

El promedio global de aciertos en toda la muestra ($m = 0{,}564$) es superior al 50 %, indicador de una dificultad media del instrumento, como corresponde a este tipo de pruebas, y que confirma la adecuación de la demanda cognitiva del test al desarrollo cognitivo de los estudiantes destinatarios.

Tabla 2.

Proporción de aciertos medios en las 18 cuestiones evaluadas con el test de pensamiento crítico para el grado 5.º de primaria (RdP_EP5; n = 146)

			Índices de discriminación	
	Índices de dificultad (m)	Desviación Estándar (DE)	Correlación 18 ítems	Correlación 16 ítems
CLA_Areas1	0,500	0,502	0,408	0,419
CLA_Areas2	0,719	0,451	0,272	0,288
CLA_Areas3	0,658	0,476	0,310	0,309
CLA_Areas4	0,644	0,481	0,416	0,438
PRO_Perro_Niño	0,719	0,451	0,277	0,303
PRO_Perro_Niña	0,788	0,410	0,326	0,344
PRO_Perro_Padre	0,774	0,420	0,340	0,367

			Índices de discriminación	
	Índices de dificultad (m)	Desviación Estándar (DE)	Correlación 18 ítems	Correlación 16 ítems
DEC_CUBITO	0,329	0,471	0,238	0,230
CLA_imagen_dif	0,466	0,501	0,383	0,409
DEC_balanza1	0,363	0,483	0,185	0,175
DEC_balanza2	0,363	0,483	0,396	0,394
DEC_balanza3	0,329	0,471	0,371	0,350
PRO_Casa 1	0,603	0,491	0,284	0,264
PRO_Casa 2	0,664	0,474	0,308	0,265
PRO_Casa 3	0,603	0,491	0,323	0,315
PRO_Casa 4	0,774	0,420	0,360	-
RL8_lapices	0,377	0,486	0,337	0,300
RL17_librosSara	0,479	0,501	0,105	-

Fuente: elaboración propia.

Los promedios de respuestas correctas en las 18 cuestiones muestran una distribución equilibrada entre preguntas fáciles y difíciles (Tabla 2). La gran mayoría de cuestiones (13) presenta un índice de aciertos intermedio (0,70 - 0,30), una minoría (5) es fácil (m > 0,700) y no hay ninguna difícil (m < 0,30) ni muy fácil (m > 0,80). Estos índices de dificultad hacen del test una prueba que supone un reto interesante y motivador para los estudiantes.

Las correlaciones entre los ítems miden el índice de discriminación de las 18 preguntas y del modelo final de 16 ítems, entre los cuales no hay diferencias notables (Tabla 2). La mayoría de las 18 cuestiones (10) alcanza un índice de discriminación bueno (0,30 – 0,40), una minoría (4) moderado (0,20 – 0,30), dos muy buenos (> 0,40) y dos bajos (< 0,20).

Las puntuaciones totales, obtenidas al sumar los aciertos de cada estudiante, muestran una distribución entre una puntuación mínima de 3 puntos y una máxima de 18 (M = 10,15; DE = 3,62), que satisface los criterios de normalidad (Shapiro-Wilk, p = .083) y homogeneidad de varianzas (Levene, p = 0,995).

Análisis factorial confirmatorio del modelo del test

El método de residuos mínimos, la rotación *oblimin* y el análisis paralelo (AP) de extracción de factores se aplicaron en sucesivos AFE para explorar el número óptimo de dimensiones en las distintas soluciones factoriales. Los parámetros de factibilidad del AFE presentan un valor apropiado de esfericidad, aunque el valor de KMO es moderado (Tabla 3).

La exploración de las soluciones AFE con uno, dos o tres factores arroja parámetros alejados de los criterios de bondad de ajuste y estructuras carentes de sentido teórico, por lo que estos modelos se descartaron.

La solución AFE con una estructura de cuatro factores presenta parámetros que no satisfacen del todo los criterios de ajuste aceptable, aunque se aproximan. La Tabla 3 (Modelo 1) muestra que el valor de χ^2 es significativo, RMSEA > 0,05 y TLI < 0,90, aunque otros índices (χ^2/gl < 5 y SRMR < 0,08) pueden considerarse adecuados. El factor de mayor peso en este modelo agrupa ocho ítems, cuya base son los cinco ítems teóricamente asignados a *clasificación*, un ítem de *decisiones* y los dos ítems de *razonamiento lógico* (ambos con cargas bajas). Los otros tres factores aparecen definidos, sin cargas cruzadas, con cargas altas y sentido teórico. Uno de los factores está compuesto por tres ítems de la destreza *decisiones* (situación Balanza), y los otros dos resultan del desdoblamiento de los siete ítems de *resolución de problemas*: uno formado por los cuatro ítems de la situación *casa* y otro por los tres ítems de la situación *perro*.

Tabla 3.

Índices generales de bondad de ajuste para las distintos modelos factoriales analizados mediante análisis factorial exploratorio y confirmatorio

		Modelo 2	Modelo 3	Modelo 4
	Modelo 1	**M_5F_16I**	**M_4F_16I**	**M_4F_16I_final**
Análisis	AFE	AFC	AFC	AFC
N.° de factores	4	5	4	4
Método	Residuos mínimos y rotación Oblimin	Análisis Paralelo	Análisis Paralelo	Análisis Paralelo
Bartlett (Esfericidad)	< 0,001			
KMO (Muestreo)	0,656			
Ítems	18	16	16	16
χ^2 (mínimo)	132	121	133	120
gl	87	94	98	98
Sig. (p > .05)	0,001	0,031	,010	0,064
χ^2/gl (< 5)	1,52	1,29	1,36	1,22
SRMR (≤ .08)		0,0608	0,0663	0,0620
RMSEA (p ≤ .05)	0,0592	0,0445	0,0496	0,0393
(IC del 95 %) Inferior	0,0378	0,0144	0,0251	0,00
(IC del 95 %) Superior	0,0797	0,0659	0,0697	0,0614
CFI (≥ .95)		0,943	0,926	0,954
TLI (≥ .95)	0,836	0,927	0,909	0,943
AIC (mínimo)		2746	2750	2722
BIC (mínimo)	-302	2919	2912	2883

Fuente: elaboración propia.

Análisis factorial confirmatorio

La solución AFE anterior, con cuatro factores, se toma como referencia para refinar otras soluciones de cuatro y cinco factores mediante AFC, considerando pautas adicionales como la disminución de los valores sucesivos de χ^2, AIC y BIC y la significación de los estimadores estandarizados, con el fin de hallar la solución con los mejores índices de ajuste.

Se exploró primero una solución AFC de cinco factores (no expuesta en la Tabla 3), cuya estructura podría ser teóricamente interpretable, aunque los parámetros de ajuste se

alejan de la aceptabilidad. Además, los estimadores estandarizados no significativos de dos ítems (DEC_Balanza1 y PRO_Casa3) sugirieron que su eliminación podría mejorar el modelo. La solución obtenida tras aplicar dicha eliminación corresponde al Modelo 2 (M_5F_16I), con 16 ítems, que presenta estimadores estandarizados significativos en todos ellos y una mejora generalizada de los índices de ajuste respecto al modelo anterior: menores valores de χ^2, AIC y BIC y valores adecuados de varios criterios (χ^2/gl < 5, SRMR < 0,08). Sin embargo, los valores de RMSEA, χ^2, AIC y BIC no alcanzan los puntos de corte óptimos, aunque algunos se aproximan (Tabla 3).

El análisis de residuos e índices de modificación de este modelo no permitió alcanzar mejoras sustanciales, por lo que se exploraron soluciones de cuatro factores.

La primera solución AFC de cuatro factores muestra una estructura teóricamente coherente, pero sus parámetros de ajuste siguen algo alejados de los puntos de corte. Como en el caso anterior, dos ítems (DEC_Balanza1 y PRO_Casa4) presentaron estimadores estandarizados no significativos, y su eliminación generó una solución de 16 ítems (Modelo 3, M_4F_16I) con mejoras respecto al modelo previo: valores de χ^2, AIC y BIC inferiores, y varios índices adecuados (SRMR = 0,0663; RMSEA = 0,0496; χ^2/gl = 1,36). No obstante, los índices CFI (0,926) y TLI (0,909) no alcanzaron el punto de corte excelente (0,95), aunque mejoraron.

Tabla 4.

Estructura final del test RdP_EP5 formada por cuatro factores empíricos, ajustada mediante análisis factorial confirmatorio, y cuyos ítems obtienen todos estimadores estándar significativos (p < 0,001)

Factores (Omega)	Ítem	Estimador	EE	Z	p	Estimador Estándar
Factor 1	PRO_Perro_Niño	0,432	0,0350	12,34	< 0,001	0,961
	PRO_Perro_Niña	0,257	0,0324	7,91	< 0,001	0,628
	PRO_Perro_Padre	0,297	0,0334	8,89	< 0,001	0,709
Factor 2	PRO_Casa 1	0,398	0,0407	9,78	< 0,001	0,813
	PRO_Casa 2	0,272	0,0398	6,84	< 0,001	0,576
	PRO_Casa 3	0,393	0,0407	9,67	< 0,001	0,803
Factor 3	CLA_Areas1	0,306	0,0466	6,57	< 0,001	0,613
	CLA_Areas2	0,175	0,0437	4,01	< 0,001	0,389
	CLA_Areas3	0,184	0,0453	4,05	< 0,001	0,387
	CLA_Areas4	0,270	0,0442	6,11	< 0,001	0,564
	CLA_imagen_dif	0,269	0,0460	5,85	< 0,001	0,540
	DEC_CUBITO	0,209	0,0450	4,65	< 0,001	0,445
	RL8_lapices	0,165	0,0473	3,49	< 0,001	0,341
Factor 4	DEC_balanza2	0,371	0,0521	7,12	< 0,001	0,771
	DEC_balanza3	0,307	0,0480	6,39	< 0,001	0,653
	DEC_balanza1	0,195	0,0457	4,28	< 0,001	0,406

Fuente: elaboración propia.

Asimismo, esta solución presenta un ítem (RL17_librosSara) con estimador estandarizado no significativo. Al eliminarlo, se obtuvo un nuevo modelo de 15 ítems que mejoró el ajuste. Sin embargo, al reponer el ítem DEC_Balanza1, eliminado en una etapa anterior, no solo no se redujo la bondad de ajuste, sino que se mejoraron los indicadores, ya que los estimadores estandarizados de todos los ítems resultaron significativos ($p < 0{,}001$) y la coherencia teórica de la estructura final, con 16 ítems, aumentó (Tabla 4).

En síntesis, este modelo final (Modelo 4, M_4F_16I_final) mejora los índices del modelo anterior (Modelo 3 M_4F_16I): los valores de χ^2 (120), AIC (2722) y BIC (2883) disminuyen, y los demás índices se ajustan excelentemente a los puntos de corte (χ^2 no significativa, aunque próxima, $p = 0{,}064$; SRMR = 0,0620; RMSEA = .0393; $\chi^2/gl = 1{,}22$; CFI = 0,954; TLI = 0,943).

La interpretación teórica de esta estructura es coherente con la construcción inicial del test (Tablas 4 y 5). Los factores primero y segundo derivan del desdoblamiento de los siete ítems iniciales de la destreza *resolución de problemas* (PRO_Perro y PRO_Casa) en dos factores de tres ítems cada uno, manteniendo su consistencia interna. El tercer factor del modelo final (siete ítems) está compuesto por los cinco ítems asociados teóricamente a la destreza *clasificación*, junto con un ítem de *decisiones* (DEC_CUBITO) y otro de *razonamiento lógico* (RL8_lapices). Cabe señalar, no obstante, la carga factorial baja de algún ítem de este tercer factor, lo que sugiere futuras mejoras (Tabla 5). Finalmente, el cuarto factor está integrado por los tres ítems de la situación *balanza*, vinculados a la destreza *decisiones*.

La varianza explicada por este modelo final de cuatro factores es satisfactoria (40,8 %), y se observa, además, que los cuatro factores contribuyen de manera bastante homogénea a dicha varianza común (Tabla 5).

Tabla 5.

Estructura de cargas en los cuatro factores del modelo 4 final (ajustado) del test RdP_EP5

	Factor				
	1	2	3	4	Unicidad
PRO_Perro_Niño	0,940				0,130
PRO_Perro_Padre	0,676				0,490
PRO_Perro_Niña	0,631				0,580
PRO_Casa 3		0,809			0,357
PRO_Casa 1		0,807			0,346
PRO_Casa 2		0,548			0,629
DEC_CUBITO			0,649		0,575
CLA_Areas4			0,492		0,679
CLA_Areas1			0,483		0,666
CLA_imagen_dif			0,461		0,711
RL8_lapices			0,380		0,822
CLA_Areas3			0,307		0,841
CLA_Areas2			0,243		0,879
DEC_balanza2				0,795	0,373
DEC_balanza3				0,623	0,580
DEC_balanza1				0,405	0,811
Consistencia interna (omega)	0,815	0,778	0,665	0,650	0,725
Varianza explicada (%)	12,4	10,92	9,52	8,02	40,8

Nota: el método de extracción 'Residuo mínimo' se usó combinado con una rotación *oblimin*. Las cargas bajas (< 0,20) se han eliminado.

Fuente: elaboración propia.

A partir de esta estructura del modelo final de 16 ítems (M_4F_16I_final), se calculó la consistencia interna del test, obteniéndose un valor global bueno (omega de McDonald = 0,725). La consistencia interna de cada uno de los cuatro factores del modelo final (M_4F_16I_final) alcanza valores aceptables del parámetro omega (0,650 a 0,815), a pesar del número reducido de ítems (tres) en tres de los factores (Tabla 4), lo cual constituye un indicio moderado de validez convergente de los factores.

Análisis correlacional: validez discriminante y por criterio externo

La Tabla 6 presenta las correlaciones y covarianzas medias entre los cuatro factores. Todas las correlaciones son apreciables, con excepción de las observadas entre el factor 2 y los factores 1 y 3.

Tabla 6.

Correlaciones y covarianzas entre los cuatro factores empíricos del modelo 4 final (ajustado) del test RdP_EP5

	Varianza media extraída	Correlaciones		
Factores		**2**	**3**	**4**
1	0,580	0,003	0,228	0,106
2	0,535		0,025	0,253
3	0,201			0,292
4	0,395			

Fuente: elaboración propia.

Asimismo, las varianzas medias extraídas de los factores son mayores que las respectivas correlaciones al cuadrado entre ellos, lo que constituye un indicio importante de validez discriminante del test según el criterio de Fornell y Larcker (1981).

Para una muestra restringida de estudiantes (n = 95) se obtuvieron las calificaciones escolares de la asignatura de ciencias, tomadas como criterio externo para contrastar la validez del test RdP_EP5. Se calculó la correlación de Pearson entre las calificaciones finales del curso en ciencias y las puntuaciones del test. El coeficiente de correlación de Pearson entre la calificación escolar final y la puntuación total del test (18 ítems) fue positivo y significativo ($r = 0,335$; $p < 0,001$), resultado similar al obtenido con la versión reducida de 16 ítems ($r = 0,325$; $p < 0,001$). Ambos resultados indican que la varianza común compartida entre ambas variables (PC y calificaciones) supera el 10 %.

Conclusiones y discusión

Este estudio aporta evidencias sobre la validez y la fiabilidad del test *Retos de pensamiento* (RdP_EP5) para evaluar el PC, diseñado como una prueba libre de

cultura (independiente del currículo escolar) y adaptada a la etapa evolutiva y de aprendizaje de los estudiantes de quinto grado de primaria (10-11 años). La estructura factorial del test RdP_EP5 presenta cuatro factores, con siete, tres, tres y tres ítems, respectivamente, cuya interpretación teórica es consistente con los resultados empíricos; los índices de bondad de ajuste son excelentes y la consistencia interna del test completo y de los factores es aceptable.

El estudio sigue las recomendaciones para el desarrollo de tests orientadas a establecer las propiedades psicométricas del RdP_EP5 (Ferrando *et al.*, 2022; Muñiz y Fonseca-Pedrero, 2019). En particular, los procedimientos empleados para explorar el número de dimensiones del modelo factorial óptimo mediante la exclusión progresiva de ítems con parámetros inadecuados reproducen el análisis de las nuevas cargas en las estructuras factoriales posteriores tras eliminar un ítem deficiente (Lloret-Segura *et al.*, 2014). Así, el modelo empírico que ofrece el mejor ajuste es aquel que proporciona una solución con cuatro factores y 16 ítems (sin cargas cruzadas entre factores), aunque algunas cargas son bajas (< 0,40). Los parámetros de bondad de ajuste del modelo son excelentes, y los índices de fiabilidad tanto del RdP_EP5 (0,725) como de cada uno de los cuatro factores empíricos identificados alcanzan puntuaciones aceptables, considerando que algunos factores están compuestos por pocos ítems (3).

Además, la interpretación de la solución factorial ajustada es congruente con el modelo teórico inicial del instrumento en relación con las destrezas *clasificación*, *resolución de problemas*, *toma de decisiones* y *razonamiento lógico*. El factor 4 se corresponde con la destreza *toma de decisiones*; el factor más amplio (3) se asocia con la destreza *clasificación* (que incorpora dos ítems nuevos), y los ítems de la destreza *resolución de problemas* se desdoblan en dos factores independientes (1 y 2), no correlacionados entre sí.

La evidencia de validación se fundamenta en la credibilidad y especialización académica de las fuentes de los ítems iniciales y en el pilotaje previo de numerosos ítems, cuya selección permitió construir esta versión del RdP_EP5. Además, el coeficiente de correlación de Pearson, positivo y significativo, entre la puntuación del test y un criterio externo (las calificaciones escolares), junto con la mayor covarianza media entre factores respecto a sus correlaciones, avalan empíricamente la validez del instrumento.

La validación psicométrica del RdP_EP5 (y su versión para sexto grado, RdP_EP6), junto con su simplicidad de aplicación y puntuación, permite su uso directo y práctico en educación primaria. Esta herramienta resulta útil y funcional para visibilizar el pensamiento y su progreso en las aulas de primaria y en la investigación educativa. Los educadores e investigadores pueden diagnosticar y evaluar de manera fácil y fiable el PC de los estudiantes, lo que posibilita comprobar la eficacia de la enseñanza del PC, ámbito en el que la escasez de evaluaciones fiables deja una incógnita sobre la efectividad de las prácticas pedagógicas, especialmente en las asignaturas donde el PC es esencial (Colom *et al.*, 2014; Saiz, 2017).

La importancia educativa del test RdP_EP5 radica en el impacto transversal del PC sobre todos los aprendizajes escolares, en la creciente expansión de su enseñanza en las escuelas —con la consecuente necesidad de evaluar los resultados— y en la falta de instrumentos de evaluación adaptados a los estudiantes jóvenes y apropiados para su uso en el aula (Aktoprak y Hursen, 2022; Ennis, 2018; Wang y Chen, 2024). Asimismo, el RdP_EP5 permite monitorear el progreso de las competencias de PC en estudios longitudinales del sistema

educativo, de modo que se pueda evaluar tanto el impacto de las competencias en el aprendizaje como el efecto del aprendizaje en las competencias, ambos aspectos cruciales para la calidad de la educación (Hattie, 2012; OECD, 2018; Unesco, 2016).

Una limitación del instrumento RdP_EP5 proviene de la obvia restricción de su diseño a cuatro competencias, aunque estas sean apropiadas para los estudiantes de quinto grado de educación primaria. Por otro lado, el proceso de validación muestra algunos parámetros con valores modestos —como el índice KMO moderado y las cargas pequeñas de ciertos ítems—, lo que sugiere acciones futuras de mejora del test. Otras limitaciones son el bajo número de ítems (3) en tres de los factores del modelo empírico y en la destreza inicial de *razonamiento lógico* (2), lo que puede haber contribuido a la desaparición de esta dimensión en el modelo final. También sería conveniente aplicar el test a otras muestras para comprobar la estabilidad de la bondad de ajuste en diferentes contextos. Estas consideraciones abren posibilidades de mejora del instrumento, ampliándolo con nuevos ítems y aplicándolo a nuevas poblaciones, a fin de fortalecer y consolidar su validación (Ferrando *et al.*, 2022; Muñiz y Fonseca-Pedrero, 2019).

En resumen, los resultados demuestran la validez y fiabilidad del test, sustentadas en una interpretación teórica coherente con los resultados empíricos, índices de bondad de ajuste excelentes y una consistencia interna aceptable tanto en el test completo como en los cuatro factores empíricos (formados por siete, tres, tres y tres ítems). El RdP_EP5 es, por tanto, una herramienta sólida y sencilla para evaluar la competencia cognitiva en PC de los estudiantes de quinto grado de primaria, aplicable al monitoreo del progreso de distintos grupos, la evaluación de la efectividad de los programas de enseñanza del PC y la comparación de resultados entre diversas estrategias o programas educativos.

Referencias

Aktoprak, A. y Hursen, C. (2022). A Bibliometric and Content Analysis of Critical Thinking in Primary Education. *Thinking Skills and Creativity*, (44). https://doi.org/10.1016/j.tsc.2022.101029

Almerich, G., Suárez-Rodríguez, J., Díaz-García, I. y Orellana, N. (2020). Estructura de las competencias del siglo XXI en alumnado del ámbito educativo. Factores personales influyentes. *Educación XX1*, *23*(1), 45-74. https://doi.org/10.5944/educxx1.23853

Bailin, S., Case, R., Coombs, J. y Daniels, L. (1999). Conceptualizing Critical Thinking. *Journal of Curriculum Studies*, *31*(3), 285-302. https://www.tandfonline.com/toc/tcus20/31/3

Boateng, G., Neilands, T., Frongillo, E., Melgar-Quiñonez, H. y Young, S. (2018). Best Practices for Developing and Validating Scales for Health, Social, and

Behavioral Research: A Primer. *Frontiers in Public Health*, (6). https://doi.org/10.3389/fpubh.2018.00149

Colom, R., García-Moriyón, F., Magro, C. y Morilla, E. (2014). The long-term Impact of Philosophy for Children: A Longitudinal Study (Preliminary Results). *Analytic Teaching and Philosophical Praxis*, *35*(1), 50-56. https://journal.viterbo.edu/index.php/atpp

Dwyer, C., Hogan, M. y Stewart, I. (2014). An Integrated Critical Thinking Framework for the 21st Century. *Thinking Skills and Creativity*, (12), 43-52. https://doi.org/10.1016/j.tsc.2013.12.004

Ennis, R. y Chattin, G. (2018). *An Annotated List of Critical Thinking Tests*. https://criticalthinking.net/wp-content/uploads/2024/04/An-Annotated-List-of-English-Language-Critical-Thinking-Tests.pdf

Ennis, R. (2018). Critical Thinking across the Curriculum: A Vision. *Topoi*, (37), 165-184. https://doi.org/10.1007/s11245-016-9401-4

Ennis, R. y Millman, J. (2005a). *Cornell Critical Thinking Test Level X*. The Critical Thinking Company. https://www.criticalthinking.com/cornell-critical-thinking-test-level-x.html

Ennis, R. y Millman, J. (2005b). *Cornell Critical Thinking Test Level Z*. The Critical Thinking Company. https://www.criticalthinking.com/cornell-critical-thinking-test-level-z.html

European Union. (2014). *Key Competence Development in School Education in Europe: KeyCoNet's Review of the Literature: A Summary*. Key Competence Network y European Schoolnet. http://keyconet.eun.org

Facione, P. (1990). *Critical Thinking: A Statement of Expert Consensus for Purposes of Educational Assessment and Instruction*. Insight Assessment.

Facione, P., Blohm, S., Howard, K. y Giancarlo, C. (1998). *California Critical Thinking Skills Test: Manual*. California Academic Press.

Ferrando, P. y Lorenzo-Seva, U. (2018). Assessing the Quality and Appropriateness of Factor Solutions and Factor Score Estimates in Exploratory Item Factor Analysis. *Educational and Psychological Measurement*, *78*(5), 762-780. https://doi.org/10.1177/0013164417719308

Ferrando, P., Lorenzo-Seva, U., Hernández-Dorado, A. y Muñiz, J. (2022). Decalogue for the Factor Analysis of Test Items. *Psicothema*, *34*(1), 7-17. https://doi.org/10.7334/psicothema2021.456

Fisher, A. (2009). *Critical Thinking: An Introduction*. Cambridge University Press.

Fisher, A. (2021). What Critical Thinking is. En J. Blair (ed.), *Studies in Critical Thinking* (2.ª ed., pp. 7-26). University of Windsor. https://scholar.uwindsor.ca/philosophybooks/8

Fornell, C. y Larcker, D. (1981). Evaluating Structural Equation Models with Unobservable Variables and Measurement Error. *Journal of Marketing Research*, *18*(1), 39-50. https://doi.org/10.2307/3151312

Fullan, M. y Scott, G. (2014). *Education* PLUS. Collaborative Impact SPC.

Gelerstein, D., Río, R. del., Nussbaum, M., Chiuminatto, P. y López, X. (2016). Designing and Implementing a Test for Measuring Critical Thinking in Primary School. *Thinking Skills and Creativity*, (20), 40-49. https://doi.org/10.1016/j.tsc.2016.02.002

Halpern, D. (2003). *Thought and Knowledge: An Introduction to Critical Thinking* (4.ª ed.). Lawrence Erlbaum.

Halpern, D. (2010). *Halpern Critical Thinking Assessment*. SCHUHFRIED. http://www.schuhfried.com/vienna-test-system-vts/all-tests-from-a-z/test/hcta-halpern-critical-thinking-assessment-1

Hattie, J. (2009). *Visible Learning: A Synthesis of over 800 meta-analyses Relating to Achievement*. Routledge. https://www.routledge.com/Visible-Learning-A-Synthesis-of-Over-800-Meta-Analyses-Relating-to-Achievement/Hattie/p/book/9780415476188

Hattie, J. (2012). *Visible Learning for Teachers: Maximizing Impact on Learning*. Routledge. https://www.routledge.com/Visible-Learning-for-Teachers-Maximizing-Impact-on-Learning/Hattie/p/book/9780415690157

International Society for Technology Education. (2003). *National Educational Technology Standards for Teachers: Preparing Teachers to Use Technology*. Autor. https://iste.org/es/standards/iste-standards-for-teachers

Lai, E. (2011). Critical Thinking: A Literature Review. *Pearson Research Reports*, (6), 1-49. https://dl.icdst.org/pdfs/files/0d632bad5f600c0564b4297ba1f8d352.pdf

Lloret-Segura, S., Ferreres-Traver, A., Hernández-Baeza, A. y Tomás-Marco, I. (2014). El análisis factorial exploratorio de los ítems: una guía práctica, revisada y actualizada. *Anales de Psicología*, *30*(3), 1151-1169. https://doi.org/10.6018/analesps.30.3.199361

Lopes, J., Silva, H. y Morais, E. (2018). Teste de pensamento crítico para estudantes dos ensinos básico e secundário. *Revista de Estudios e Investigación en Psicología y Educación*, *5*(2), 82-91. https://doi.org/10.17979/reipe.2018.5.2.3339

Lorenzo-Seva, U. y Ferrando, P. (2019). Robust Promin: A Method for Diagonally Weighted Factor Rotation. *Liberabit: Revista Peruana de Psicología*, (25), 99-106. https://doi.org/10.24265/liberabit.2019.v25n1.08

Manassero-Mas, M. y Vázquez-Alonso, Á. (2020a). Evaluación de destrezas de pensamiento crítico: validación de instrumentos libres de cultura. *Tecné, Episteme y Didaxis,* TED, (47), 15-32. https://doi.org/10.17227/ted.num47-9801

Manassero-Mas, M. y Vázquez-Alonso, Á. (2020b). Las destrezas de pensamiento y las calificaciones escolares en educación secundaria: validación de un instrumento de evaluación libre de cultura. *Tecné, Episteme y Didaxis,* TED, (48), 33-54. https://doi.org/10.17227/ted.num48-12375

Manassero-Mas, M. y Vázquez-Alonso, Á. (2023). Evaluación de las destrezas del pensamiento crítico: un diagnóstico de los estudiantes de primaria. *Revista Evaluar*, *23*(2), 40-56. https://doi.org/10.35670/1667-4545.v23.n2.42069

Manassero-Mas, M. y Vázquez-Alonso, Á. (2024). Visibilizar las destrezas de pensamiento en educación primaria: desarrollo psicométrico de un instrumento de evaluación. *Bordón: Revista de Pedagogía*, *76*(1), 119-139. https://doi.org/10.13042/BORDON.2024.95702

Manassero-Mas, M., Moreno-Salvo, A. y Vázquez-Alonso, Á. (2022). Development of an Instrument to Assess Young People's Attitudes toward Critical Thinking. *Thinking Skills and Creativity*, (45). https://doi.org/10.1016/j.tsc.2022.101100

Mapeala, R. y Siew, N. (2015). The Development and Validation of a Test of Science Critical Thinking for Fifth Graders. *SpringerPlus*, *4*(1). https://doi.org/10.1186/s40064-015-1535-0

Muñiz, J. y Fonseca-Pedrero, E. (2019). Ten Steps for Test Development. *Psicothema*, *31*(1), 7-16. https://doi.org/10.7334/psicothema2018.291

National Research Council. (2012). *Education for Life and Work: Developing Transferable Knowledge and Skills in the 21st Century*. The National Academies Press. https://doi.org/10.17226/13398

O'Hare, L. y McGuinness, C. (2015). The Validity of Critical Thinking Tests for Predicting Degree Performance: A Longitudinal Study. *International Journal of Educational Research*, (72), 162-172. https://doi.org/10.1016/j.ijer.2015.06.004

Organisation for Economic Co-operation and Development. (2018). *Future of education and skills*. Autor. https://www.oecd.org/en/topics/future-of-education-and-skills.html

Piaget, J. e Inhelder, B. (1997). *Psicología del niño*. Morata.

Saiz, C. (2017). *Critical Thinking and Change*. Pirámide.

Shayer, M. y Adey, P. (Eds.). (2002). *Learning Intelligence: Cognitive Acceleration across the Curriculum from 5 to 15 Years*. Open University Press.

Szőköl, I., Szarka, K. y Hargaš, J. (2022). The Functions of Educational Evaluation. R&E-SOURCE. https://doi.org/10.53349/RESOURCE.2022.IS24.A1112

Tremblay, K. (2013). OECD Assessment of Higher Education Learning Outcomes (AHELO). En S. Blömeke, O. Zlatkin-Troitschanskaia, C. Kuhn y J. Fege (eds.), *Modeling and Measuring Competencies in Higher Education* (pp. 113-126). Sense Publishers. https://doi.org/10.1007/978-94-6091-867-4_8

Unesco. (2016). *Education 2030: Incheon Declaration and Framework for Action for the Implementation of Sustainable Development Goal 4*. Autor. https://unesdoc.unesco.org/ark:/48223/pf0000245656

Valenzuela, J. (2008). Habilidades de pensamiento y aprendizaje profundo. *Revista Iberoamericana de Educación*, (46). https://doi.org/10.35362/rie4671914

Vázquez-Alonso, Á. y Manassero-Mas, M. (2018). Una taxonomía de las destrezas de pensamiento: una herramienta clave para la alfabetización científica. *Tecné, Episteme y Didaxis,* TED, (extraordinario), 1-7. https://revistas.upn.edu.co/index.php/TED/article/view/9189

Vincent-Lancrin, S., González-Sancho, C., Bouckaert, M., De Luca, F., Fernández-Barrera, M., Jacotin, G., Urgel, J. y Vidal, Q. (2019). *Fostering Students' Creativity and Critical Thinking*. OECD. https://doi.org/10.1787/62212c37-en

Wang, X. y Chen, J. (2024). The Investigation of Critical Thinking Disposition among Chinese Primary and Middle School Students. *Thinking Skills and Creativity*, (51). https://doi.org/10.1016/j.tsc.2023.101444

Watson, G. y Glaser, E. (2002). *Watson-Glaser Critical Thinking Appraisal-II Form E*. Pearson.

World Economic Forum. (2021). *These Are the Top 10 Job Skills of Tomorrow*. Autor. https://www.weforum.org/agenda/2020/10/top-10-work-skills-of-tomorrow-how-long-it-takes-to-learn-them