VISIBILIZAR LAS DESTREZAS DE PENSAMIENTO EN EDUCACIÓN PRIMARIA: DESARROLLO PSICOMÉTRICO DE UN INSTRUMENTO DE EVALUACIÓN

Making thinking skills visible in elementary education: psychometric development of an evaluation tool

MARÍA ANTONIA MANASSERO MAS Y ÁNGEL VÁZQUEZ ALONSO
Universidad de las Islas Baleares (España)

DOI: 10.13042/Bordon.2024.95702
Fecha de recepción: 21/07/2022 • Fecha de aceptación: 25/01/2024
Autora de contacto / *Corresponding autor*: María Antonia Manassero. E-mail: ma.manassero@uib.es

Cómo citar este artículo: 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

INTRODUCCIÓN. Las competencias del siglo XXI incluyen siempre el pensamiento crítico (PC) y proyectan una demanda creciente de innovación educativa, porque el PC no es un contenido usual en la educación escolar. La carencia de instrumentos de evaluación de PC para primaria dificulta su visibilidad y educación, y justifica el objetivo de este estudio: validar un instrumento de evaluación del PC libre de cultura para educación primaria y establecer sus propiedades psicométricas. **MÉTODO**. Las prescripciones habituales para desarrollos empíricos de test son seguidas en un proceso con dos formas del test aplicadas a dos muestras de estudiantes de sexto grado de educación primaria y refinamientos estadísticos entre ambas. Se aplican métodos correlacionales de análisis factorial exploratorio y confirmatorio para determinar la fiabilidad y validez de la prueba, construida con cinco destrezas teóricas apropiadas para la educación primaria: predicción, comparación, clasificación, resolución de problemas y razonamiento lógico. **RESULTADOS**. Los resultados describen los estadísticos de ítems y destrezas, y confirman una estructura empírica de un instrumento final de 29 ítems y cuatro factores empíricos (predicción-confirmación, clasificación ampliada, resolución de problemas y nuevo razonamiento). La parsimoniosa interpretación de los factores junto con los parámetros apropiados de bondad del ajuste (GFI .962) y de la fiabilidad de los factores y la prueba final (.966) apoyan la validez psicométrica de la prueba para evaluar el pensamiento en educación primaria. **DISCUSIÓN**. Las propiedades psicométricas de validez, bondad de ajuste y fiabilidad del instrumento prueban su utilidad para la visibilidad, la evaluación y la investigación del pensamiento en primaria. Además, el carácter unidimensional de los cuatro factores empíricos justifica las aplicaciones independientes de cada uno de ellos. Finalmente, algunas limitaciones psicométricas sugieren potenciales líneas prospectivas de futuros desarrollos y aplicaciones para continuar mejorando la calidad del instrumento.

Palabras clave: *Pensamiento crítico, Test culturalmente justo, Validez, Fiabilidad, Enseñanza primaria.*

Introducción

La enseñanza de las competencias del siglo XXI proyecta una demanda global continuamente creciente sobre la educación en las sociedades del conocimiento para afrontar desafíos tales como la globalización, la progresiva influencia científica y tecnológica, la acelerada innovación digital e informativa, la emergencia ecológica y sus impactos en la vida personal, laboral y social (Almerich *et al.*, 2020). Esas competencias engloban destrezas digitales y cognitivas, y estas últimas distinguen entre destrezas blandas (interpersonales) y duras (pensamiento de alto nivel), que también se denominan pensamiento crítico (PC).

Las múltiples habilidades cognitivas de alto nivel que forman el PC proceden de un desarrollo complejo de las categorías superiores de la taxonomía de Bloom (analizar, juzgar y crear) (Krathwohl, 2002). Algunas de estas destrezas más regularmente citadas en la literatura son, por ejemplo, argumentación, análisis, interpretación, creatividad, resolución de problemas, investigación, toma de decisiones, metacognición, así como diversas combinaciones de estas y otras.

Las destrezas del PC se consideran aspectos clave para el aprendizaje significativo y profundo, muy sensible al dominio de destrezas de pensamiento (Valenzuela, 2008). La educación del PC entronca con los pioneros estudios de Piaget en el siglo XX (Piaget y Inhelder, 1997) y los programas de aceleración cognitiva (Shayer y Adey, 2002), que han demostrado empíricamente su significativo impacto sobre el aprendizaje. El metaanálisis del aprendizaje visible de Hattie (2009, 2012) informa de que el tamaño del efecto sobre el aprendizaje de los programas piagetianos es muy alto (d=1.28) y el impacto de otras variables de PC (estrategias metacognitivas, creatividad, resolución de problemas, etc.) también es alto (d>.40).

Además, las destrezas de PC se consideran clave para los requerimientos de los puestos de trabajo en 2025 (Foro Económico Mundial, 2021) y esenciales para el éxito de las personas en la era de la información (Tremblay *et al.*, 2012). Por ello, múltiples instituciones y expertos apoyan la educación del PC y sus habilidades específicas asociadas (European Union, 2014; Fullan y Scott, 2014; International Society for Technology Education, 2003; National Research Council, 2012; OECD, 2018; UNESCO, 2015).

Todas estas propuestas convergen en resaltar que las destrezas de pensamiento no solo son una necesidad social, sino que siguen siendo un factor clave del aprendizaje (Moral, 2008). Ambos factores justifican la atención educativa hacia ellas, que resulta innovadora por su escasez actual, y que este estudio afronta desde una perspectiva educativa y evaluadora.

El pensamiento crítico

La investigación sobre PC ha desarrollado tres líneas básicas en el marco de la psicología cognitiva: conceptualización, enseñanza y evaluación. Sin embargo, el desarrollo de cada una ha sido desigual (Saiz, 2017).

El PC suele conceptualizarse como un tipo de pensamiento que intenta superar las tendencias naturales del pensamiento al error, la falacia y el sesgo (egocentrismo y sociocentrismo), gracias

al dominio consciente y diestro de múltiples destrezas cognitivas de alto nivel y la adhesión a disposiciones actitudinales y a estándares de calidad adecuados (Paul y Nosich, 1993). A pesar de su amplio desarrollo, la conceptualización adolece de un consenso real entre los especialistas, probablemente por las diferencias en el contexto, finalidad y objetivo de cada definición, que se hace patente en el uso de una multiplicidad de conceptos y términos que introducen una gran diversidad en el campo, como se puntualiza en los párrafos siguientes.

La conceptualización de PC propuesta por Ennis (2018), como pensamiento reflexivo y razonable centrado en decidir qué creer o hacer, y su desarrollo ampliado en disposiciones y habilidades que intervienen en esas decisiones, son muy citados. Para coordinar esta diversidad, un panel de expertos de la American Philosophical Association (APA) acordó una definición de PC como un juicio deliberado y autorregulado para un objetivo específico, que emplea las destrezas interpretación, análisis, evaluación e inferencia, basadas en evidencia, conceptos, métodos, criterios y contextos para elaborar el juicio (APA, 1990; Facione, 1990).

Como alternativa a la falta de consenso, algunos investigadores optan por definir PC por extensión, especificando las destrezas constitutivas, aunque tampoco en esta línea existe consenso (Fisher, 2009). Por ejemplo, la definición del panel APA propone ya seis destrezas (interpretación, análisis, evaluación, inferencia, interpretación, juicio y autorregulación), y, en el otro extremo, el plan nacional para la evaluación del PC propone una extensa lista de 88 elementos de PC agrupadas en cuatro dominios (Paul y Nosich, 1993).

Para paliar estas discrepancias, dos taxonomías desarrolladas recientemente presentan un marco teórico integrador que organiza el PC en cuatro dimensiones con grandes coincidencias entre ellas. Manassero-Mas y Vázquez-Alonso (2019) proponen cuatro dimensiones básicas (creatividad, razonamiento y argumentación, procesos complejos y evaluación y juicio), que contienen múltiples categorías y subcategorías (pensamiento deductivo, inductivo, abductivo y estadístico; resolución de problemas y toma de decisiones; supuestos, estándares, disposiciones). En la misma línea, Fisher (2021) ha organizado las habilidades de PC también en cuatro grupos básicos: interpretación, análisis, evaluación y autorregulación.

Paralelamente, los especialistas alcanzan consenso en otros aspectos importantes, como que los juicios de PC deben satisfacer exigentes normas y estándares de calidad, adecuación y precisión, para lograr ser un PC válido, y, en particular, para superar la lista de falacias que identifican formas de PC usuales, aunque inválidas (Bailin *et al.*, 1999).

En suma, aunque la literatura muestra diferencias entre expertos, aquí se utilizará el término PC para describir el constructo general, formado por múltiples habilidades de pensamiento y otros conceptos asociados (disposiciones actitudinales), y donde la normatividad inherente al PC es una base crucial para su evaluación.

La evaluación del pensamiento crítico

Una hipótesis implícita y generalmente asumida en la literatura es que la educación puede mejorar el pensamiento. Por ello, hace décadas que múltiples programas con variadas orientaciones y prácticas tratan de enseñar a pensar (Follmann *et al.*, 2018; Saiz, 2017; Swartz *et al.*, 2013). Sin

embargo, los programas que acreditan sus efectos empíricos con estudios de evaluación son la excepción más que la regla (Saiz, 2017). El programa de filosofía para niños de Lipman ha sido repetidamente evaluado (Colom *et al.*, 2014), mientras otros, como el aprendizaje basado en pensamiento (Swartz *et al.*, 2013, solo ocasionalmente, y otros, como el programa de razonamiento (Walton y Macagno, 2015), carecen de evaluaciones.

La declaración de los expertos APA (Facione, 1998) recomendó complementar la enseñanza del PC con su evaluación frecuente y explícita, tanto de forma diagnóstica como sumativa (recomendación 13), usando instrumentos de evaluación con validez de contenido y de constructo, fiables y equitativos (recomendación 12), hoy rasgos obvios en la construcción de todo test (Muñiz y Fonseca-Pedrero, 1999). Ennis (2018) justifica la necesidad de evaluar el PC con las razones siguientes: diagnosticar el nivel del alumnado, retroalimentar el progreso, motivar a aprender PC, informar a los docentes sobre su enseñanza, investigar el PC, asesorar la elección de estudios y estimular a las instituciones educativas para informar sus resultados.

La evaluación del PC es una necesidad y un apoyo significativo para la mejora de su enseñanza, pero requiere la construcción de instrumentos de evaluación apropiados para lograr medidas válidas y fiables. Un resumen de estos instrumentos se ha presentado y sistematizado en otro lugar (Manassero-Mas y Vázquez-Alonso, 2019; Ennis y Chattin, 2018). Sus diferentes estructuras, formatos y rasgos psicométricos han generado una amplia literatura crítica. La mayoría de ellos (Facione *et al.*, 1998; Halpern, 2010; Rivas y Saiz, 2012; Watson y Glaser, 2002) se concentran en evaluar unas pocas destrezas del PC, aunque alguno es más amplio (Madison, 2004).

La gran mayoría de los instrumentos de evaluación del PC se dirigen a adultos o estudiantes universitarios, mientras apenas hay pruebas específicas para estudiantes más jóvenes, aunque las pruebas de Cornell, denominadas X, Y, Z, son parcialmente adaptables a los jóvenes (Ennis y Millman, 2005a, 2005b) y otras propuestas requieren aún mayor consolidación psicométrica (Lopes *et al.*, 2018).

En suma, la creciente importancia del PC en la educación a través de las competencias del siglo XXI y la escasa atención a la evaluación del PC en los niveles educativos tempranos con estudiantes más jóvenes justifican la necesidad de desarrollar una prueba para la evaluación de PC en jóvenes, centrada en destrezas específicas apropiadas y en hacer visible el pensamiento en la práctica educativa de los niveles tempranos.

El objetivo de este estudio es cubrir este vacío, validando un instrumento para diagnosticar destrezas de PC en jóvenes de educación primaria. El instrumento evalúa una destreza de cada una de las cuatro dimensiones (creatividad, razonamiento, procesos complejos y juicio) de la taxonomía del PC elaborada por Manassero-Mas y Vázquez-Alonso (2019). Además, el test usa cuestiones cuya demanda cognitiva sea adecuada para primaria, culturalmente justa y sin sesgos. El estudio investiga la relación del instrumento con el aprendizaje (representado por las calificaciones escolares, como criterios externos de validez empírica), y aplica métodos de análisis factorial exploratorio y confirmatorio para apoyar la validez y la fiabilidad psicométricas.

Método

Este estudio continúa un estudio previo donde algunos estudiantes de sexto grado de primaria respondieron un banco de ítems sobre destrezas de PC (Manassero-Mas y Vázquez-Alonso, 2020a, 2020b). A partir de estos resultados se aplican las recomendaciones usuales en el desarrollo de pruebas, para eliminar cuestiones inadecuadas, adaptar otras y adicionar nuevas (Muñiz y Fonseca-Pedrero, 2019), y se construye un cuestionario con 39 ítems, que se aplica a una pequeña muestra piloto (etapa inicial). Estas respuestas se analizan nuevamente para tomar decisiones de mejora (añadir, eliminar, reformular y reasignar cuestiones), que producen una segunda forma del instrumento con 35 cuestiones, que se aplica a una nueva muestra (etapa final) y se somete a nuevos análisis factoriales exploratorios (AFE) y confirmatorios (AFC).

Participantes

Los estudiantes han sido seleccionados por pertenecer a escuelas interesadas en la educación del PC, conformando una muestra de conveniencia, aunque social y demográficamente diversa. Todos participan en este estudio por grupos naturales de clase completos en sexto curso de educación primaria.

La prueba de la etapa inicial se aplicó en 2019 a 82 estudiantes (37 hombres y 45 mujeres) con edades comprendidas entre 10 y 13 años (moda 11 años), en tres grupos naturales de tres escuelas diferentes, situadas en el centro (dos) y un barrio periférico de la misma ciudad.

La prueba final se aplicó en 2020 a 435 estudiantes (196 hombres y 239 mujeres) entre 10 y 13 años (moda 11 años), que forman 23 grupos naturales de diez escuelas diferentes, públicas (4) y concertadas (6), situadas en ciudades y poblaciones pequeñas de tres regiones distintas.

Instrumento

Sendos instrumentos, denominados "retos de pensamiento" (RdP_EP6), se han diseñado y aplicado en la etapa inicial y final, para medir destrezas de PC. Las destrezas fueron diseñadas por su adaptación al nivel cognitivo de los estudiantes participantes y por el interés específico en ellas de un centro participante; estas destrezas y las dimensiones de la taxonomía del PC (Manassero-Mas y Vázquez-Alonso, 2019) a las que pertenecen son las siguientes: predicción y razonamiento lógico (dimensión razonamiento), comparación (dimensión creatividad), clasificación (dimensión evaluación) y resolución de problemas (dimensión procesos complejos).

Las cuestiones de cada destreza fueron seleccionadas mediante un análisis minucioso de los materiales de evaluación citados en la introducción según los siguientes criterios: facilidad de lectura y comprensión, concordancia de la demanda cognitiva de cada cuestión con la destreza asignada y con el desarrollo cognitivo de los estudiantes y planteamiento de un desafío motivador e interesante para los estudiantes (tabla 1).

Tabla 1. Tabla de especificaciones de las dos pruebas aplicadas (RdP_EP6) en este estudio para evaluar destrezas de pensamiento en sexto curso de educación primaria EP6

Destrezas de pensamiento	Fuente	Información	Número de cuestiones	
			Inicial(39)	Final(35)
Predicción	Ennis y Millman, 2005a	Verbal	9	7
Comparación		Verbal	12	6
Clasificación	Elaboración propia*	Figuras	6	7
Resolución de problemas	Halpern (2010)	Verbal	8	8
Resolución de problemas	Elaboración propia*	Figuras	4	4
Razonamiento lógico	Ennis y Millman, 2005b	Verbal		3

*Fuente: * Inspirados por materiales abiertos https://www.criticalthinking.com.*

Las cuestiones plantean diversos escenarios y situaciones informativas, sobre los que se hacen una o varias preguntas, que plantean retos de pensamiento auténticos y motivadores para los estudiantes, y cuya demanda cognitiva está ajustada a la destreza que representan y al nivel evolutivo de los estudiantes. Además, sus contenidos se han diseñado sin connotaciones culturales porque no están relacionados ni anclados en conocimientos curriculares de las materias escolares, de manera que lograr la respuesta correcta no requiere conocimientos previos.

Los formatos de respuesta combinan la opción múltiple (mayoritaria), Likert (1-9) y respuestas cortas, que permiten una evaluación estandarizada, rápida, válida y fiable de cada destreza y facilitan el establecimiento de líneas base de diagnóstico para comparar investigaciones, programas y metodologías de enseñanza.

Procedimiento de recogida y análisis de datos

Los dos instrumentos, inicial y final, fueron aplicados a los participantes dentro de su grupo de clase por su profesorado siguiendo las mismas directrices estandarizadas comunes, utilizando dispositivos digitales y sin límite de tiempo (usualmente un periodo de clase) y planteados como actividades regladas ordinarias de evaluación del aprendizaje, para incentivar el esfuerzo y motivación de los estudiantes.

Las respuestas correctas reciben un punto, las incorrectas cero puntos y no se aplican correcciones por respuestas al azar. La puntuación de cada destreza es la suma de los aciertos logrados en las preguntas que la forman y la puntuación global es la suma de los aciertos totales (se considera una estimación del PC global de los estudiantes, con base en las destrezas componentes).

La validez de construcción se basa, por un lado, en la validez de las fuentes: pruebas de PC publicadas (Ennis y Millman, 2005a, 2005b; Halpern, 2010) y publicaciones especializadas en PC para las cuestiones de elaboración propia (https://www.criticalthinking.com). Por otro, se centra en el escrutinio de las cuestiones seleccionadas y el acuerdo profesional de los investigadores, con base en el mejor ajuste ítem-destreza y demanda cognitiva-nivel evolutivo de los estudiantes.

La validez de contenido se verifica con las calificaciones escolares de una muestra parcial, suministradas por los centros escolares participantes.

Los datos se procesaron con SPSS (25) y el programa Factor, que aplica un método robusto de mínimos cuadrados no ponderados (RULS), basado en correlaciones tetracóricas, apropiadas para las puntuaciones dicotómicas de cuestiones y para los AFE y AFC que extraen factores con RULS y rotación Promin. La fiabilidad, valorada mediante el índice esperado a posteriori (EAP), y otros estadísticos confirmatorios son obtenidos de los programas mencionados (Ferrando y Lorenzo-Seva, 2017, 2018; Lorenzo-Seva y Ferrando, 2019).

Resultados

Los descriptores estadísticos de los ítems de las dos pruebas en las dos etapas del estudio, obtenidos a partir las respuestas de los estudiantes están resumidos en la tabla 2.

Tabla 2. Proporción de aciertos medios (y desviación estándar) para las cuestiones evaluadas con el instrumento de PC para el grado 6 (RdP_EP6) en el estudio inicial (39 ítems; n=82; RdP_EP6_39) y el estudio final (35 ítems; n=435; RdP_EP6_35).

Etapa inicial (prueba RdP_EP6_39)		Etapa final (prueba RdP_EP6_35)	
Variables*	**Promedio aciertos (0-1) (desviación estándar)**	**Variables**	**Promedio aciertos (0-1) (desviación estándar)**
PREDIC1	.512(.503)	PREDIC1	.426(.245)
PREDIC2	.439(.499)	PREDIC2	.528(.249)
PREDIC3	.378(.488)	PREDIC3	.442(.247)
PREDIC4	.329(.473)	PREDIC4	.546(.248)
PREDIC5	.671(.473)	**	
PREDIC6	.744(.439)	**	
PREDIC7	.671(.473)	PREDIC7	.682(.217)
PREDIC8	.280(.452)	PREDIC8	.438(.246)
PREDIC9	.463(.502)	PREDIC9	.590(.242)
COMPA1	.756(.432)	**	
COMPA2	.671(.473)	**	
COMPA3	.061(.241)	**	
COMPA4	.293(.458)	COMPA4	.297(.209)
COMPA5	.573(.498)	COMPA5	.606(.239)
COMPA6	.317(.468)	COMPA6	.486(.250)
COMPA7	.500(.503)	COMPA7	.624(.235)
COMPA8	.537(.502)	**	
COMPA9	.598(.493)	**	
COMPA10	.634(.485)	COMPA10	.581(.243)
COMPA11	.488(.503)	COMPA11	.346(.226)

>>

Tabla 2. Proporción de aciertos medios (y desviación estándar) para las cuestiones evaluadas con el instrumento de PC para el grado 6 (RdP_EP6) en el estudio inicial (39 ítems; n=82; RdP_EP6_39) y el estudio final (35 ítems; n=435; RdP_EP6_35). (cont.)

Etapa inicial (prueba RdP_EP6_39)		Etapa final (prueba RdP_EP6_35)	
Variables*	**Promedio aciertos (0-1) (desviación estándar)**	**Variables**	**Promedio aciertos (0-1) (desviación estándar)**
COMPA12	.110(.315)	**	
		CLASIF0***	.588(.242)
CLASIF1	.317(.468)	CLASIF1	.491(.250)
CLASIF2	.341(.477)	CLASIF2	.523(.249)
CLASIF3	.671(.473)	CLASIF3	.802(.159)
CLASIF4	.512(.503)	CLASIF4	.703(.209)
CLASIF5	.598(.493)	CLASIF5	.673(.220)
CLASIF6	.402(.493)	CLASIF6	.535(.249)
PROBL1	.646(.481)	PROBL1	.599(.240)
PROBL2	.780(.416)	PROBL2	.869(.114)
PROBL3	.646(.481)	PROBL3	.691(.213)
PROBL4	.598(.493)	PROBL4	.544(.248)
PROBL5	.207(.408)	PROBL5	.334(.222)
PROBL6	.858(.356)	PROBL6	.855(.124)
PROBL7	.683(.468)	PROBL7	.730(.197)
PROBL8	.646(.481)	PROBL8	.705(.208)
PROBL9	.146(.356)	PROBL9	.270(.197)
PROBL10	.281(.452)	PROBL10	.415(.243)
PROBL11	.415(.496)	PROBL11	.537(.249)
PROBL12	.220(.416)	PROBL12	.406(.241)
		RAZ1***	.445(.247)
		RAZ2***	.539(.248)
		RAZ3***	.343(.225)

** La columna variables representa la tabla de especificaciones de la prueba.*

*** Ítems del instrumento inicial eliminados en el instrumento final.*

**** Ítems nuevos añadidos en el instrumento final como consecuencia del proceso de validación realizado sobre los resultados del test de la etapa inicial.*

Etapa inicial

En la etapa inicial, el promedio global de aciertos de todos los ítems es próximo al 50% (promedio=.4869), y confirma la dificultad media del instrumento, como corresponde a este tipo de pruebas (tabla 2). Además, la distribución es equilibrada entre cuestiones fáciles y difíciles; la gran mayoría de cuestiones (29) logran un promedio de respuestas correctas intermedio (.70-.30), una minoría de cuestiones (4) son muy fáciles (promedio>.70) y otra minoría de cuestiones (6) son muy difíciles (promedio<.30).

Validez respecto a criterio externo: las calificaciones escolares

Uno de los argumentos más generalizados en favor de la educación de las destrezas de PC es su impacto transversal en el aprendizaje (Hattie, 2009, 2012; Shayer y Adey, 2002; Valenzuela, 2008), con independencia del debate entre el contexto general y específico de su enseñanza. Para confirmar empíricamente la relación mutua entre destrezas de pensamiento y aprendizaje y la validación de contenido del instrumento RdP_EP6 por un criterio externo, se analizan las correlaciones entre las puntuaciones de las destrezas y las calificaciones escolares en una submuestra de estudiantes.

La distribución de puntuaciones máximas y mínimas de las calificaciones escolares finales obtenidas por los estudiantes en las asignaturas curriculares del sexto grado de educación primaria presenta cierta asimetría entre asignaturas: la máxima es alcanzada en todas las asignaturas, pero las mínimas se distribuyen irregularmente (tabla 3).

Tabla 3. Estadística descriptiva de las calificaciones en las asignaturas escolares obtenidas por los participantes en el estudio inicial

Asignaturas	Mínimo	Máximo	Promedio	Desviación estándar
Ciencias Naturales	2	10	6.68	1.81
Ciencias Sociales	2	10	6.70	1.80
Educación Artística	2	10	6.95	1.83
Educación Física	5	10	7.65	1.65
Lengua Castellana	6	10	8.04	0.87
Lengua Catalana	4	10	7.30	1.45
Lengua Inglesa	4	10	7.09	1.40
Matemáticas	3	10	7.09	1.63
Religión	4	10	7.24	1.48

Las correlaciones entre PC y asignaturas (tabla 4) son todas positivas y mayoritariamente significativas. Entre las dos variables globalizadoras de ambos constructos (global de PC y nota media de asignaturas) exhiben los valores más altos y significativos. La puntuación global de PC muestra las correlaciones más altas con matemáticas y religión; las correlaciones más bajas, aunque significativas, se establecen con lengua castellana.

Desde la perspectiva de las destrezas, comparación y resolución de problemas tienen las correlaciones más altas y significativas con las asignaturas (solo una excepción respectivamente, lengua castellana y educación física). En el otro extremo, la destreza predicción es la más baja, pues solo correlaciona significativamente con matemáticas. Clasificación no correlaciona con las calificaciones de educación artística, educación física y lengua castellana.

Desde la perspectiva de las asignaturas, las correlaciones son más heterogéneas. Por un lado, un grupo de asignaturas (ciencias sociales, lengua catalana y matemáticas) establecen correlaciones significativas con las cuatro destrezas de PC; por otro lado, educación física y lengua castellana solo establecen relaciones significativas con una destreza de pensamiento y el resto mantienen correlaciones significativas con tres destrezas.

Tabla 4. Correlaciones entre puntuaciones de las destrezas de PC y global de pensamiento con las calificaciones de las asignaturas y la nota media de las calificaciones (n=82)

Calificaciones	Destrezas de pensamiento (RdP_EP6_39)				
	Predicción	Comparación	Clasificación	Problema	Global
Ciencias Naturales	.208	.427**	.255*	.424**	.562**
Ciencias Sociales	.220*	.420**	.249*	.431**	.564**
Educación Artística	.255*	.352**	.158	.338**	.466**
Educación Física	.211	.371**	.175	.193	.394**
Lengua Castellana	.078	.117	.098	.303**	.264*
Lengua Catalana	.254*	.394**	.332**	.426**	.596**
Lengua Inglesa	.178	.431**	.307**	.428**	.576**
Matemáticas	.297**	.377**	.350**	.535**	.665**
Religión	.177	.368**	.327**	.566**	.626**
Nota media	.258*	.420**	.290**	.459**	.608**

** Correlación significativa en el nivel .05 (bilateral).*

*** Correlación significativa en el nivel .01 (bilateral).*

En suma, las correlaciones entre destrezas de PC (sin connotaciones culturales) y calificaciones escolares (centradas en contenidos curriculares) indican que ambos constructos diferentes correlacionan mayoritaria y significativamente entre sí, siendo especialmente altas con la puntuación global de PC. Estos resultados apoyan la validez del instrumento respecto a un criterio externo (calificaciones) y demuestran la importancia general del PC para los aprendizajes escolares que numerosos estudios postulan o apoyan.

A pesar de la pequeña muestra, se ha realizado un AFE del instrumento inicial basado en correlaciones tetracóricas que apunta fortalezas y debilidades. La adecuación de muestreo (Kaiser-Meyer-Olkin KMO) y la proporción de varianza total explicada (29%) son bajas y doce autovalores son negativos. Sin embargo, una solución de cuatro factores empíricos ajusta parcialmente los cuatro factores postulados teóricamente, pues cada factor incluye un núcleo de cuestiones pertenecientes a la destreza teórica que representa; este ajuste es más robusto en los factores representativos de clasificación y problemas, y menos en predicción y comparación. No obstante, una decena de ítems de las tres últimas destrezas mencionadas muestran cargas bajas o negativas o aparecen sin ubicación en ningún factor.

Mediante AFC se ha contrastado separadamente la naturaleza unidimensional de las cuatro destrezas teóricas, obteniendo parámetros de bondad de ajuste bajos, aunque las cuatro destrezas obtienen valores aceptables de fiabilidad EAP (>.70). Las destrezas predicción y comparación presentan valores bajos del índice KMO y de la proporción de varianza explicada (24%-19% frente a 49% de clasificación) y algunas cuestiones tienen cargas negativas en predicción (4), comparación (3) y resolución de problemas (1).

Estos resultados sugieren revisar el instrumento para mejorarlo, eliminando ítems deficientes (8), adicionando cuestiones nuevas (4), señaladas con asteriscos en la cuarta columna (variables) de la prueba final (tabla 2) y ampliando la muestra en la etapa final.

Etapa final

La revisión del instrumento inicial RdP_EP6_39 produce una nueva versión RdP_EP6_35 que se aplicó a una muestra mayor en la etapa final; los promedios y desviaciones de los ítems de RdP_EP6_35 están resumidos en la mitad derecha de la tabla 2. Los resultados de aciertos muestran una distribución equilibrada entre cuestiones fáciles y difíciles, con la gran mayoría de cuestiones (27) que logran un promedio de respuestas correctas intermedio (entre .70 y .30), una minoría de cuestiones (6) muy fáciles (aciertos>.70) y otra minoría de cuestiones (2) muy difíciles (aciertos<.30). El promedio de aciertos global (.548) confirma la dificultad intermedia de RdP_EP6_35, como se espera de pruebas de este tipo.

El AFC de RdP_EP6_35 con el método robusto RULS y correlaciones tetracóricas obtiene valores favorables de los parámetros KMO (.89341) y probabilidad (.000001) y produce un conjunto de autovalores, sobre los cuales el análisis paralelo factorial de rango mínimo sugiere soluciones de tres, cuatro (aconsejada atendiendo al promedio de los autovalores) o cinco factores empíricos que se analizan a continuación (tabla 5).

Tabla 5. Parámetros estadísticos de la bondad de ajuste robusta confirmatoria de los modelos de factores contrastados para la prueba final

Parámetros estadísticos	**Modelos contrastados**				
Número de ítems	35	35	35	30º	29ª
Factores extraídos	3	5	4	4	4
Ji-cuadrado	1046.9	980.0	851.1	532.8	491.4
Ji-cuadrado (p)	.00001	.00001	.00001	.00001	.00001
RMSEA*	.051	.038	.044	.039	.039
NNFI**	.944	.969	.958	.975	.977
CFI***	.954	.978	.967	.982	.983
GFI****	.925	.951	.940	.959	.962
RMSR*****	.076	.062	.068	.063	.062
WRMSR******	.0470	.0385	.0429	.0384	.0374
Varianza explicada	.3441	.4442	.3974	.4491	.4629
Fiabilidad (EAP)					
Factor1	.823	.719	.935	.982	.847
Factor2	.932	.944	.981	.947	.954
Factor3	.952	.756	.821	.852	.990
Factor4		.986	.947	.694	.677
Factor5		.997			
Total					.966

**Root Mean Square Error of Approximation*

***Normed Fit Index*

****Comparative Fit Index*

*****Goodness of Fit Index*

******Root Mean Square of Residuals (aceptable próximo a .048)*

*******WRMSR Weighthed Root Mean Square of Residuals (aceptables< 1.0)*

ºÍtems eliminados PREDIC7, COMPA4, COMPA11, PROBL5, PROBL7

ªÍtems eliminados PREDIC7, COMPA4, COMPA11, PROBL4, PROBL5, PROBL7

Los parámetros del modelo de tres factores empíricos sugieren su descarte porque la heterogeneidad de los ítems en cada factor, las cargas cruzadas entre factores y las cargas negativas de algunos impiden una interpretación razonable de la naturaleza de cada factor.

El modelo de cinco factores presenta un factor principal consolidado, pero la estructura de los restantes cuatro factores es tan compleja que no permite una interpretación global, sencilla y coherente de los factores.

Además, el análisis comparado de los parámetros confirmatorios entre los tres modelos con 35 ítems (tabla 5) es favorable al modelo de cuatro factores y los cuatro factores de este modelo admiten una interpretación sencilla y razonable: tres factores corresponden principalmente a las destrezas teóricas postuladas desde el inicio (clasificación, resolución de problemas y razonamiento), y, el cuarto factor se considera una fusión de ítems de predicción y comparación.

Sin embargo, el AFE, AFC y las propiedades psicométricas de cada factor empírico muestran que los factores predicción y resolución de problemas tienen valores más bajos de fiabilidad y otros parámetros y algunos ítems muestran cargas factoriales muy bajas o negativas en todos los factores. Estos resultados sugieren refinar el ajuste del modelo de cuatro factores eliminando los ítems deficientes.

En consecuencia, un nuevo AFC con rotación Promin compara dos nuevos modelos de cuatro factores, que prescinden de cinco y seis ítems disfuncionales (tabla 5, dos columnas últimas). El modelo de cuatro factores con 29 ítems ofrece los mejores parámetros de bondad de ajuste. El nuevo cómputo de la matriz de cargas factoriales con 6 ítems menos visualiza la constitución de cuatro factores empíricos que permiten una interpretación simple y coherente entre la estructura y los contenidos de los ítems que conforman los factores empíricos (tabla 6).

Los ítems asignados a los cuatro factores empíricos reflejan las destrezas teóricas de partida con matices. El factor clasificación ampliado muestra cargas positivas y relevantes de los ítems originales de la destreza clasificación; además, incorpora cuatro ítems, procedentes de predicción y comparación. El factor resolución de problemas queda limitado a los cinco ítems verbales supervivientes de la eliminación previa de ítems defectuosos y pierde los ítems figurativos de la destreza original.

Los factores obtenidos son oblicuos y correlacionados, por lo que las cargas son coeficientes de regresión (y pueden ser mayores que 1)

Los factores nuevo razonamiento y predicción-comparación son el resultado de sendas fusiones de ítems provenientes de diferentes destrezas (tabla 6). Nuevo razonamiento integra en un nuevo factor los tres ítems de razonamiento con los cuatro ítems figurativos asignados a resolución de problemas. Predicción-comparación fusiona los ítems de las destrezas predicción y comparación originales (con excepción de los tres ítems que cargan significativamente en clasificación).

Tabla 6. Matriz de cargas factoriales con rotación oblicua Promin del instrumento reducido RdP_EP6_29 (29 ítems)

Ítems	Resolución de problemas	Clasificación ampliada	Nuevo razonamiento	Predicción-comparación
Predic1			.306	.426
Predic2		.412		
Predic3				.555
Predic4		.324		
Predic8				.197
Predic9				.403
Compa5				.214
Compa6		.217		
Compa7		.392		
Compa10				.511
Clasif0		.419		
Clasif1		1.054		
Clasif2		.610		
Clasif3	.427	.139		
Clasif4		.483		
Clasif5		.651		
Clasif6	-.347	1.107		
Probl1	.469			
Probl2	.904			
Probl3	.152	.337		
Probl6	.639			
Probl8	.359			
Probl9		.232	.178	
Probl10			.986	
Probl11			.831	
Probl12			1.050	
Raz1			.200	
Raz2			.146	
Raz3	-.465		.340	

Nota: suprimidas las cargas inferiores a .30 (exceptuando las cargas de variables asignadas a los cuatro factores).

Análisis confirmatorio de la unidimensionalidad de los cuatro factores empíricos

Cada uno de los cuatro grupos de ítems que conforman los cuatro factores empíricos anteriores se someten separadamente a un análisis RULS para verificar independientemente su carácter unidimensional. Los resultados globales obtenidos en los cuatro factores presentan buenos índices de bondad de ajuste, varianza explicada y fiabilidad, pero también sugieren algunas mejoras (tabla 7).

Los análisis paralelos basados en los AF de rango mínimo confirman para los cuatro factores un modelo de dimensión única. El parámetro MIREAL presenta también valores aceptables (< .30), con una sola moderada excepción, que permiten considerar estos cuatro factores como unidimensionales. En consecuencia, sus puntuaciones miden válida y fiablemente las destrezas representadas y pueden aplicarse con independencia del resto de la prueba.

Tabla 7. Análisis factorial confirmatorio de la unidimensionalidad de los cuatro factores empíricos del modelo reducido RdP_EP6_29 apoyado por el AFC previo

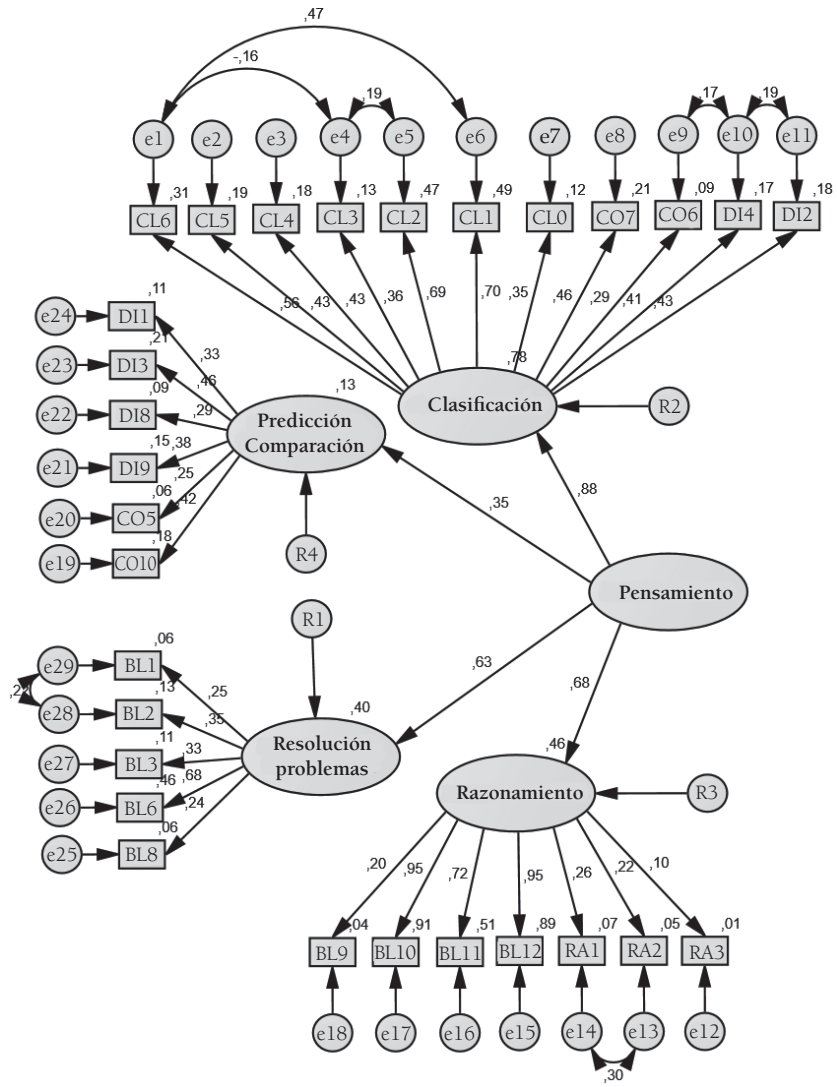
Estadísticos de AFC	Predicción-comparación		Clasificación ampliada		Resolución de problemas		Nuevo razonamiento	
	Ítem	Carga	Ítem	Carga	Ítem	Carga	Ítem	Carga
	Predic1	.434	Clasif0	.447	Probl1	.493	Probl9	.337
	Predic3	.632	Clasif1	.933	Probl2	.659	Probl10	.965
	Predic8	.258	Clasif2	.806	Probl3	.317	Probl11	.913
	Predic9	.500	Clasif3	.444	Probl6	.830	Probl12	.983
	Compa5	.148	Clasif4	.565	Probl8	.307	Raz1	.415
	Compa10	.371	Clasif5	.562			Raz2	.405
			Clasif6	.753			Raz3	.128
			Predic2	.541				
			Predic4	.493				
			Compa6	.148				
			Compa7	.542				
Kaiser-Meyer-Olkin	.688773		.73640		.64798		.60713	
Bartlett (Sig.)	.000000		.000010		.000000		.000000	
Varianza explicada	.28577		.44545		.42334		.48947	
Unidimensional								
MIREAL*	.231		.271		.321		.287	
Bondad ajuste								
RMSEA**	.022		.085		.092		.082	
Ji Cuadrado	16.999		114.529		23.259		54.388	
Ji Cuadrado (P)	.0263098		.000010		.000000		.000001	
NNFI ***	.984		.955		.909		.965	
CFI****	.989		.965		.955		.976	
GFI*****	.985		.958		.970		.977	
Fiabilidad (EAP)	.631		.926		.779		.979	

El modelo de factores empíricos (tabla 7) para el cuestionario RdP_EP6_29 (tabla 5) se valida con un análisis de diagramas mediante ecuaciones estructurales (programa AMOS), cuyo resultado incluye covarianzas en algunos errores de observables (figura 1). Los parámetros de bondad de ajuste del diagrama mejoran los valores de otros diagramas y son favorables: los valores de

ajuste absoluto, tanto dependientes de la muestra (Chi-cuadrado=592.394, p=.000), como los independientes de la muestra (Chi-cuadrado normal=1.619; RMSEA=.038), el ajuste incremental (NFI=.804, TLI=.904, CFI=.913) y el ajuste de la parsimonia (PRATIO=. 901 e índice AIC mejor que otros diagramas).

El factor más extenso y figurativo (clasificación) muestra la correlación más alta con la variable latente de pensamiento, mientras la más baja corresponde al factor predicción-comparación (verbal). En general, las cuestiones figurativas tienden a tener correlaciones superiores a las verbales y todos los factores tienen algunos observables con cargas bajas.

Figura 1. Coeficientes de regresión estandarizados entre las variables latentes y los observables del análisis factorial confirmatorio para el modelo final, que muestra covarianzas entre algunos errores de observables (AMOS 26)



Discusión y conclusiones

Este estudio tiene dos objetivos principales: validar psicométricamente un instrumento en lengua española para la evaluación del pensamiento crítico (RdP_EP6) y establecer sus propiedades y dimensiones. El instrumento es libre de cultura y adaptado al desarrollo evolutivo de los estudiantes de sexto grado de primaria, y cuya necesidad está justificada por la carencia de instrumentos de evaluación apropiados para los más jóvenes y la creciente extensión de la educación del PC (Ennis y Chattin, 2018).

El proceso en dos etapas se adapta a las prescripciones generales del desarrollo de tests (Muñiz y Fonseca-Pedrero, 2019) y parte de una prueba inicial (39 ítems) con cuatro destrezas teóricas (predicción, comparación, clasificación y resolución de problemas), que alcanza una buena validación por criterio externo (calificaciones escolares), pero muestra deficiencias psicométricas que sugieren su revisión y producen una prueba revisada (35 ítems) que se aplica en una segunda etapa a una muestra más amplia. Sucesivos análisis y refinamientos mediante AFE y AFC permiten alcanzar una forma final (RdP_EP6_29) con solo 29 ítems y con una estructura simple y parsimoniosa de cuatro factores empíricos nuevos, cuyos buenos valores de bondad del ajuste apoyan la validez estructural del modelo, junto con una buena fiabilidad global (.966) y de los cuatro factores empíricos (.631 a .979).

El instrumento final RdP_EP6_29 de 29 ítems mide cuatro factores empíricos (predicción-confirmación, clasificación ampliada, resolución de problemas y nuevo razonamiento), que evalúan sendas destrezas genuinamente cognitivas de PC, a diferencia de otros instrumentos (Lopes *et al.*, 2018). La escala predicción-comparación valora la capacidad para verificar una conclusión a partir de un razonamiento inductivo (predicción) o desde el contraste entre varias afirmaciones (comparación). La escala de clasificación valora la capacidad para agrupar o separar diferentes elementos según rasgos comunes o diferenciales. La destreza de resolución de problemas mide la capacidad para valorar las mejores y peores soluciones ante situaciones problemáticas de la vida cotidiana. Finalmente, la escala de razonamiento valora la capacidad deductiva simple (silogismo simple) y compleja (cuando intervienen simultáneamente varias informaciones o conclusiones).

Las implicaciones prácticas son contribuir a hacer visible el pensamiento y su progreso en el aula de primaria y en la investigación, gracias a la sencillez de la aplicación y puntuación de la prueba elaborada. Primero, el instrumento puede ser una herramienta útil para educadores, psicólogos e investigadores para diagnosticar del PC y evaluar la eficacia de programas de intervención específicos (Colom *et al.*, 2014; Saiz, 2017). Segundo, el instrumento permite la visibilidad e identificación de las destrezas de PC y proporciona retroalimentación actualizada a estudiantes y docentes sobre su progreso y seguimiento; además, el carácter unidimensional de los cuatro factores garantiza evaluaciones independientes de cada factor, que eviten la fatiga con la prueba completa (OECD, 2018; UNESCO, 2015). Tercero, las evaluaciones con el instrumento pueden informar tanto futuras investigaciones sobre destrezas de PC como aplicarse en estudios longitudinales para evaluar el impacto de las destrezas en el aprendizaje y de este en aquellas (Hattie, 2012).

No obstante, algunos parámetros estadísticos con valores moderados suponen sendas limitaciones del test que sugieren futuros refinamientos. Así, el factor resolución de problemas muestra valores moderados de KMO, MIREAL y NNFI, nuevo razonamiento puede mejorar su KMO, el factor predicción-comparación tiene fiabilidad y varianza moderadas y varias cuestiones presentan cargas

bajas en su factor. La implementación de refinamientos y nuevas aplicaciones con nuevas muestras es un reto para superar las limitaciones apuntadas (Muñiz y Fonseca-Pedrero, 2019).

Prospectivamente, se espera que la investigación futura mejore el RdP_EP6_29 y proporcione evidencia adicional sobre su validez y fiabilidad. La aplicación a estudiantes de distintas edades ampliaría la variabilidad de respuestas, calibraría mejor la validez y fiabilidad de RdP_EP6_29 y consolidaría más la utilidad educativa de la prueba, en aspectos tales como su baremación para diferentes grupos, la relación con otras medidas cognitivas del PC y con las calificaciones escolares, así como el análisis de la validez predictiva entre ellas.

Agradecimientos

Proyecto EDU2015-64642-R (AEI/FEDER, UE), financiado por la Agencia Estatal de Investigación (AEI) y el Fondo Europeo de Desarrollo Regional (FEDER).

Referencias bibliográficas

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. R. y Daniels, L. B. (1999). Conceptualizing critical thinking. *Journal of Curriculum Studies*, *31*(3), 285-302. https://doi.org/10.1080/002202799183133

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*, 50-55. https://journal.viterbo.edu/index.php/atpp/article/view/1129

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

Ennis, R. H. y Chattin, G. S. (2018). An annotated list of critical thinking tests. http://criticalthinking.net/wp-content/uploads/2018/01/An-Annotated-List-of-English-Language-Critical-Thinking-Tests.pdf

Ennis, R. H. 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. H. 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*. http://keyconet.eun.org

Facione, P. A. (1990). *The Delphi Report. Critical Thinking: A Statement of Expert Consensus for purposes of Educational Assessment and Instruction. Executive Summary*. California Academic Press. https://www.qcc.cuny.edu/socialsciences/ppecorino/CT-Expert-Report.pdf

Facione, P. A. (1998). *Insight Assessment*. www.insightassessment.com

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

Ferrando, P. J. y Lorenzo-Seva, U. (2017). Program FACTOR at 10: Origins, development and future directions. *Psicothema*, *29*, 236-240. https://doi.org/ 10.7334/psicothema2016.304

Ferrando, P. J. 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*, 762-780. https://doi.org/10.1177/0013164417719308

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

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

Follmann, D., Mattos, K. R. C. y Güllich, R. I. da C. (2018). Estratégias de Ensino de Ciências e a Promoção de Pensamento Crítico em Portugal. *Tecné, Episteme y Didaxis* (Extra). https://revistas.pedagogica.edu.co/index.php/TED/article/view/8789

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

Halpern, D. F. (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*. International Society for Technology Education. https://iste.org/es/standards/iste-standards-for-teachers

Krathwohl, D. (2002). A Revision of Bloom's Taxonomy. *Theory into Practice*, *41*, 212-218. https://doi.org/10.1207/s15430421tip4104_2

Lopes, J., Silva, H. y Morais, E. (2018). Teste de pensamento crítico para estudantes dos ensinos básico e secundário [Critical thinking test for elementary and secondary students]. *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. J. (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

Madison, J. (2004). *James Madison Critical Thinking Course*. The Critical Thinking Co. https://www.criticalthinking.com/james-madison-critical-thinking-course.html

Manassero-Mas, M. A. y Vázquez-Alonso, A. (2019). Taxonomía de las destrezas de pensamiento: una herramienta clave para la alfabetización científica. En M. D. Maciel y E. Albrecht (orgs.), *Ciência, Tecnologia y Sociedade: Ensino, Pesquisa e Formação* (pp. 17-38). UNICSUL.

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

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

Moral Santaella, C. (2008). Aprender a pensar-aprender a aprender. Habilidades de pensamiento y aprendizaje autorregulado. *Bordón. Revista de Pedagogía*, *60*(2), 123-137. https://recyt.fecyt.es/index.php/BORDON/article/view/29019

Muñiz, J. y Fonseca-Pedrero, E. (2019). Diez pasos para la construcción de un test. *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

OECD (2018). The future of education and skills. Education 2030. http://go.uv.es/1fDpQnn

Paul, R. y Nosich, G. M. (1993). A Model for the National Assessment of Higher Order Thinking. En R. Paul (ed.), *Critical Thinking: What Every Student Needs to Survive in a Rapidly Changing World* (pp. 78-123). Foundation for Critical Thinking. https://www.criticalthinking.org/pages/a-model-for-the-national-assessment-of-higher-order-thinking/591

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

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

Swartz, R. J., Costa, A. L., Beyer, B. K., Reagan R. y Kallick, B. (2013). *El aprendizaje basado en el pensamiento*. SM.

Tremblay, K., Lalancette, D. y Roseveare, D. (2012). Assessment of Higher Education Learning Outcomes. Design and Implementation, *Feasibility Study Report*, vol. 1. http://www.oecd.org/education/skills-beyond-school/AHELOFSReportVolume1.pdf

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

Walton, D. y Macagno, F. (2015). A Classification System for Argumentation Schemes. *Argument and Computation*, *6*(3), 214-249. https://doi.org/10.1080/19462166.2015.1123772

Watson, G. y Glaser, E. M. (2002). *Watson-Glaser Critical Thinking Appraisal-II Form E*. Pearson. https://www.pearsonvue.com/phnro/wg_practice.pdf

World Economic Forum (2021). These are the top 10 job skills of tomorrow – and how long it takes to learn them. https://www.weforum.org/agenda/2020/10/top-10-work717skills-of-tomorrow-how-long-it-takes-to-learn-them/

Abstract

Making thinking skills visible in elementary education: psychometric development of an evaluation tool

INTRODUCTION. The competencies of the 21st century always include critical thinking (CT) and project a growing demand on educational innovation, because CT is not a usual content of school education. The lack of CT evaluation instruments for primary education hinders its visibility and teaching and justifies the aim of this study: to validate a cultural-free CT evaluation test for primary education and to establish its psychometric properties. METHOD. The usual prescriptions for empirical test development are followed in a process involving two test forms that are applied to two different samples of sixth grade of primary education, and statistical refinements between the two forms. Correlational exploratory and confirmatory factor analysis are applied to check test validity and reliability. By construction the test assesses five theoretical skills: prediction, comparison, classification, problem solving and logical reasoning. RESULTS. The results describe the statistics of items and skills, and confirm an empirical structure of a 29-item final test and four empirical factors (prediction-confirmation, expanded classification, problem solving and new reasoning). The parsimonious interpretation of the factors together with appropriate goodness-of-fit parameters (GFI .962) and the reliability of the factors and the final test (.966) support the psychometric validity of the test to evaluate thinking in primary education. DISCUSSION. The instrument's psychometric properties of validity, goodness-of-fit

and reliability prove its usefulness for visibility, evaluation and research of thinking in primary education. Furthermore, the unidimensional nature of the four empirical factors justifies independent applications of each of them. Finally, some psychometric limitations suggest potential prospective lines for future developments and applications to continue improving the quality of the instrument.

Keywords: *Critical thinking, Culture fair test, validity, Reliability, Elementary education.*

Résumé

Rendant visibles les habilités de pensée dans l'enseignement primaire : développement psychométrique d'un instrument d'évaluation

INTRODUCTION. Parmi les compétences du XXI siècle se trouve toujours la pensée critique (PC), à la fois qu'elle projette une demande croissante d'innovation pédagogique, car la PC n'est pas un contenu habituel de l'enseignement scolaire. La manque d'instruments d'évaluation de la PC au primaire entrave sa visibilité et son enseignement en justifiant l'intêret de cette étude: la validation d'un instrument culturellement équitable d'évaluation de la PC au primaire en établissant ses caractéristiques psychométriques. **MÉTHODE**. Les prescriptions habituelles pour le développement des tests sont adoptées au cours d'un processus avec deux formes de test appliquées à deux échantillons d'élèves de sixième année du primaire ainsi que des améliorations statistiques pour les deux formulaires. Des méthodes corrélationnelles d'analyse factorielle exploratoire et confirmatoire sont appliquées pour déterminer la fiabilité et la validité du test, construit avec cinq compétences théoriques appropriées à l'enseignement primaire : prédiction, comparaison, classification, résolution de problèmes et raisonnement logique. **RÉSULTATS**. Les résultats montrent les statistiques des items et de compétences confirmant une structure empirique d'un instrument final de 29 items et de quatre facteurs empiriques (prédiction-confirmation, classification élargie, résolution de problèmes et nouveau raisonnement). L'interprétation parcimonieuse des facteurs, les paramètres appropriés d'adéquation (GFI .962) ainsi que la fiabilité des facteurs et du test final (.966) soutiennent la validité psychométrique du test d'évaluation de la pensée critique dans l'enseignement primaire. **DISCUSSION**. Les propriétés psychométriques de validité, d'adéquation et de fiabilité de l'instrument prouvent son utilité pour la visibilité, l'évaluation et la recherche sur la pensée à l'école primaire. En outre, la nature unidimensionnelle des quatre facteurs empiriques justifie les applications indépendantes de chacun d'eux. Certaines limites psychométriques suggèrent des pistes potentielles de développements et d'applications futures pour l'amélioration de la qualité de l'instrument.

Mots-clés : *Pensée critique, Test culturellement équitable, Validité, Fiabilité, Enseignement primaire.*

Perfil profesional de los autores

María Antonia Manassero-Mas (autora de contacto)

Doctora en Psicología, Premio Extraordinario de Licenciatura y Doctorado y catedrática de Psicología Social. Ha sido directora de la Universidad Abierta para Mayores, Defensora Universitaria y directora de varios grupos de investigación. Participante en más de dos decenas de proyectos competitivos y autora de centenares de libros y capítulos de libros, artículos en revistas y comunicaciones a congresos internacionales y conferenciante invitada. Pertenece a asociaciones científicas y consejos editoriales y ejerce de revisora y evaluadora de artículos y proyectos de investigación científicos para revistas y agencias de investigación en el ámbito de la psicología y la educación.
ORCID https://orcid.org/0000-0002-7804-7779
Correo electrónico de contacto: ma.manassero@uib.es.
Dirección para la correspondencia: Facultad de Psicología, Universidad de las Islas Baleares, Edificio Margalida Comas i Camps, Carretera de Valldemossa, km 7.5. 07122 Palma, España

Ángel Vázquez-Alonso

Doctor en Educación, máster en Ciencias Físicas y licenciado en Química y Educación. Ha servido como profesor de Bachillerato, inspector de educación, director del Instituto de Evaluación de las Islas Baleares y profesor de la Universidad de las Islas Baleares. Participante en más de dos decenas de proyectos competitivos y autor de decenas de libros y capítulos de libros, dos centenares de artículos en revistas arbitradas, otras tantas publicaciones en congresos y ha impartido cursos y conferencias invitadas en decenas de eventos. Evaluador de proyectos de investigación e innovación para agencias europeas, americanas y españolas, revisor para revistas y congresos nacionales e internacionales y miembro de varias organizaciones profesionales.
ORCID https://orcid.org/0000-0001-5830-7062
Correo electrónico de contacto: angel.vazquez@uib.es.