Evaluación de destrezas de pensamiento en educación secundaria: propiedades psicométricas de un instrumento independiente de la cultura

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

- Assessment of Thinking Skills in Secondary Education: Psychometric Properties of a Culture-fair Instrument
- Avaliação das habilidades de pensamento no ensino médio: Propriedades psicométricas de um instrumento independente da cultura

Resumen

La enseñanza de las destrezas de pensamiento constituye una creciente demanda para los sistemas educativos debido a su transversalidad respecto a competencias y aprendizajes, además de su funcionalidad personal y laboral en las actuales sociedades del conocimiento. En la investigación psicológica se conceptualizan como destrezas de pensamiento crítico y se han desarrollado programas de enseñanza y evaluación, principalmente para adultos. Este artículo presenta un instrumento para evaluar destrezas de PC dirigida a jóvenes, que independiente de su cultura, desarrolla un componente figurativo importante, a diferencia de otros instrumentos centrados en aprendizajes específicos. El instrumento evalúa cuatro destrezas: deducción, asunciones, secuenciación y toma de decisiones, en estudiantes de educación secundaria (15-16 años) y su validación empírica toma las calificaciones escolares como criterio externo, compara los dos cursos de los participantes, y otros procedimientos. Los resultados muestran las características psicométricas del instrumento de diagnóstico de las destrezas de pensamiento, donde cabe resaltar la correlación significativa y alta entre las calificaciones escolares y las destrezas de pensamiento, además de la discriminación del instrumento en el diagnóstico diferencial de dos cursos sucesivos. La validez y fiabilidad de las cuestiones figurativas son mejores que las cuestiones verbales, aunque estas proceden de instrumentos estandarizados. Finalmente, se discute la importancia educativa de crear un instrumento de evaluación independiente de la cultura y un lenguaje común para la enseñanza que permitan desarrollar la enseñanza y evaluación de las destrezas de pensamiento en las aulas; además, se concretan algunas propuestas de revisión para mejorar la validez y fiabilidad del instrumento de evaluación.

Palabras clave

pensamiento crítico; evaluación de destrezas; validez; fiabilidad

* Profesora catedrática de Psicología Social, doctora en Psicología, Universidad de las Islas Baleares, Palma de Mallorca, España. ma.manassero@uib.es. Orcid: https://orcid.org/0000-0002-7804-7779.

** Profesor asociado e investigador honorífico, Doctor en educación, Centro de Estudios de Posgrado, Universidad de las Islas Baleares, Palma de Mallorca, España. angel.vazquez@uib.es. Orcid: https://orcid.org/0000-0001-5830-7062.

Fecha de recepción: 24/01/2022
Fecha de aprobación: 03/06/2022

Abstract

Teaching thinking skills is a growing demand for educational systems due to its transversal role regarding competencies and learning and its functionality for persons and jobs in current knowledge societies. In psychological research, these skills are conceptualized as critical thinking, where teaching programs and some assessment tests, mainly aimed at adults, have been created. This study displays a culture-fair critical thinking assessment instrument, which addresses young people and includes figurative components, unlike other evaluation tools that are focused on specific learning. The instrument assesses four thinking skills, deduction, assumptions, sequences, and decision-making, and aims to students of secondary education (15-16 years). Its empirical validation considers students' school marks and the comparison between the participants' two grades, as external criteria, together with other procedures. The results show the psychometric characteristics of the assessment instrument for diagnosing students thinking skills; the most important findings are the significant and high correlations between school marks and thinking skills and the instrument discrimination for the differential diagnosis of two successive educational grades. Further, the validity and reliability of the figurative items are better than those of the verbal items, despite the latter being drawn from standardized instruments. Finally, the educational importance of creating a culture fair critical thinking assessment instrument and a common language that allows the development of the teaching thinking skills and their culture-free assessment within the classroom; further, some revision proposals for improving the instrument validity and reliability are suggested.

Keywords

critical thinking; skill assessment; validity; reliability

Resumo

O ensino das habilidades do pensamento constitui uma demanda crescente para os sistemas educacionais, devido à sua transversalidade no que diz respeito a competências e aprendizagens, além da sua funcionalidade pessoal e profissional nas atuais sociedades do conhecimento. Na pesquisa psicológica conceitualizam-se como habilidades do pensamento critico e desenvolvem-se programas de ensino e avaliação, principalmente para adultos. Este artigo apresenta um instrumento para avaliação habilidades de PC dirigida a jovens, que independente da sua cultura, desenvolve um importante componente figurativo, ao contrário de outros instrumentos centrados em aprendizagens específicas. O instrumento avalia quatro habilidades: dedução, premissas, sequenciamento e tomada de decisões, em alunos do ensino médio (15 a 16 anos) e a sua validação empírica emprega as notas escolares como critério externo, a comparação entre os dois cursos dos participantes, entre outros procedimentos. Os resultados mostram as características psicométricas do instrumento do diagnóstico das habilidades de pensamento, onde o achado mais importante é a correlação significativa e elevada entre notas escolares e habilidades de pensamento, além da discriminação do instrumento no diagnóstico diferencial dos dois cursos sucessivos. A validade e confiabilidade das questões figurativas são melhores do que as questões verbais, embora estas venham de instrumentos padronizados. Por fim, discute-se a importância educacional de criar um instrumento da avaliação independente da cultura e uma linguagem comum que facilitam o ensino e a avaliação de habilidades de pensamento nas salas de aula; além disso, são feitas algumas propostas de revisão para melhorar a validade e a confiabilidade do instrumento de avaliação.

Palavras-chave

pensamento crítico; avaliação de habilidades; validade; confiabilidade

Introducción

El futuro de las actuales sociedades del conocimiento afronta numerosos desafíos, tales como el creciente impacto científico y tecnológico, la acelerada innovación digital e informativa, la globalización y la emergencia ecológica, cuyo rápido y relevante impacto sobre la vida personal y social genera múltiples demandas globales sobre los sistemas educativos, que suelen resumirse en el lema de enseñar las competencias del siglo XXI (Almerich *et al*., 2020). Esas competencias suelen dividirse en destrezas de alto nivel y destrezas digitales; las primeras contienen destrezas de pensamiento y destrezas interpersonales (comunicación, colaboración, emprendimiento, etc.). Este estudio se centra en las destrezas de pensamiento, que se suelen plantear heterogéneamente como pensamiento crítico (PC), resolución de problemas, habilidades de investigación, argumentación, resolución de problemas, análisis, interpretación, creatividad, innovación y toma de decisiones, así como expresiones que formulan diversas combinaciones de ellas (European Union, 2014; Fullan y Scott, 2014; International Society for Technology Education, 2003; National Research Council, 2012; OECD, 2018; Unesco, 2015).

Desde la perspectiva cognitiva del aprendizaje, esas múltiples destrezas se corresponden con las categorías superiores de la taxonomía de Bloom (analizar, juzgar y crear), por lo que también son denominadas destrezas de pensamiento de alto nivel (Krathwohl, 2002). Se consideran aspectos claves del aprendizaje significativo y profundo, sensibles al dominio de destrezas de pensamiento (Valenzuela, 2008) y frecuentes en las áreas de ciencias, tecnologías, ingeniería y matemáticas (STEM).

En los últimos años, diversas investigaciones didácticas sugieren el papel clave de las destrezas del pc en los aprendizajes STEM (entre otras, Ford y Yore, 2014; McDonald y McRobbie, 2012; Simonneaux, 2014; Tamayo, 2017; Tenreiro-Vieira y Vieira, 2014; Torres y Solbes, 2016; Vázquez y Manassero, 2018). Erduran y Kaya (2018) proponen claramente que los metaconceptos epistémicos de naturaleza de la ciencia siguen siendo un desafío para la educación científica, a pesar de décadas de investigación, porque exigen el dominio de las destrezas cognitivas críticas de alto nivel; el positivo impacto sobre el aprendizaje se ha justificado también por la semejanza constitutiva entre el PC y el razonamiento científico (Manassero y Vázquez, 2020a).

Los estudios pioneros de Piaget (Piaget e Inhelder, 1997), continuados por los programas de aceleración (Shayer y Adey, 2002) y otros han desarrollado las relaciones entre las destrezas de pensamiento de alto nivel y el aprendizaje. Desde la investigación educativa general se han aportado pruebas empíricas del impacto global de estas destrezas cognitivas sobre el aprendizaje. El metaanálisis del aprendizaje visible de Hattie (2009; 2012) informa que el tamaño del efecto de los programas piagetianos sobre el aprendizaje es muy alto ($d = 1{,}28$) y el impacto de otras variables de pensamiento (estrategias metacognitivas, creatividad, resolución de problemas, etc.) también es moderadamente alto ($d > ,40$).

Todas estas propuestas convergen en resaltar que las destrezas de pensamiento, ahora olvidadas en la escuela en favor de otras modas didácticas emergentes, siguen siendo un factor clave para el aprendizaje profundo y justifican la atención innovadora hacia ellas, que este estudio afronta desde una perspectiva evaluadora.

La investigación sobre pensamiento crítico

Esta investigación se ha centrado en tres líneas básicas: conceptualización, enseñanza y evaluación del constructo, aunque el desarrollo de cada una ha sido desigual (Saiz, 2017). La conceptualización ha tenido un desarrollo amplio, pero muy complejo, por la ausencia de un consenso entre los especialistas sobre algo tan básico como una definición común del PC, lo cual induce cierta complejidad terminológica del campo (por ejemplo, destrezas cognitivas, toma de decisiones, creatividad, pensamiento crítico, etc.). No obstante, el PC se suele caracterizar como una forma de pensamiento que supone un dominio consciente de múltiples destrezas cognitivas de alto nivel y la adhesión a estándares de calidad y a disposiciones permanentes a superar las tendencias a la falacia y al sesgo (egocentrismo y sociocentrismo).

La conceptualización de Ennis (2018) del PC (pensamiento reflexivo y razonable centrado en decidir qué creer o hacer) y su ampliado desarrollo en disposiciones y habilidades que intervienen en la toma de esas decisiones son muy citados. De hecho, ante la ausencia de consenso conceptual, muchos investigadores prefieren definir el PC por extensión, es decir, especificando las destrezas constituyentes (Fisher, 2009), aunque tampoco en esta línea existe consenso. El plan nacional para la evaluación del PC de Paul y Nosich (1993) propone una extensa lista de destrezas de PC (88) agrupadas en dimensiones. Sin embargo, las diferencias entre especialistas aún persisten, sean de concepto o de extensión. Para evitar una sopa de letras disfuncional, aquí se utilizará el término PC para describir esas múltiples habilidades de pensamiento y sus conceptos asociados.

La evaluación del pensamiento crítico

Asumiendo la hipótesis básica de que el pensamiento puede ser mejorado mediante programas educativos adecuados, desde hace décadas se han creado programas de enseñanza del PC con variadas orientaciones y prácticas. Sin embargo, los programas de enseñanza de PC cuyos efectos se han acreditado por estudios de evaluación empíricos son la excepción más que la regla (Saiz, 2017), pues solo el programa de filosofía para niños desarrollado por Lipman ha sido evaluado repetidamente (Colom *et al*., 2014), mientras que otros, como el aprendizaje basado en pensamiento (Swartz *et al*., 2013) solo han sido evaluados ocasionalmente, y otros carecen aún de evaluaciones, como el programa canadiense de razonamiento (Walton y Macagno, 2015). Por ello, la declaración de consenso de los expertos en PC (Facione, 1998) propone complementar su enseñanza con su evaluación y la recomendación 13 propone evaluar con frecuencia, tanto de forma diagnóstica como sumativa, todo lo cual resalta la importancia de los instrumentos de evaluación del PC.

La investigación ha desarrollado diversos instrumentos (Facione *et al.*, 1998; Halpern, 2010; Rivas y Saiz, 2012; Watson y Glaser, 2002), la mayoría de los cuales evalúan solo unas pocas destrezas, aunque uno es más amplio (Madison, 2004). Además, su población objetivo suele ser adultos y estudiantes universitarios. No hay pruebas validadas para los estudiantes jóvenes, aunque las pruebas de PC de Cornell (X, Y, Z), desarrolladas para una variedad de poblaciones, podrían aportar algunos ítems apropiados para los más jóvenes (Ennis y Millman, 2005). Este estudio trata de remediar esta carencia aportando una prueba.

Conceptualización del pensamiento crítico para la educación

La complejidad conceptual del PC es también un inconveniente para su enseñanza y evaluación en los distintos contextos educativos, ya que dificulta desarrollar un significado profesional compartido con el profesorado, que lo haga funcional en las aulas (Vincent-Lancrin *et al.*, 2019). Manassero y Vázquez (2019) han desarrollado una taxonomía de las destrezas de pensamiento con base en el análisis empírico de las múltiples conceptualizaciones de los especialistas y de las destrezas incluidas en los instrumentos de evaluación de PC, donde el constructo PC es el concepto universal y estructurante de las demás destrezas, que se agrupan en cuatro dimensiones básicas: creatividad, razonamiento y argumentación, procesos complejos y evaluación y juicio. Estas, a su vez, se dividen en categorías (pensamiento deductivo, inductivo, abductivo y estadístico; resolución de problemas y toma de decisiones; supuestos, estándares, disposiciones) y múltiples subcategorías.

Nosotros definimos el PC como: "Pensamiento creativo, claro y preciso en sus justificaciones y conclusiones, que además evalúa y juzga meticulosamente todos sus elementos".

En la misma línea, Fisher (2021) también ha organizado las habilidades del PC en cuatro grupos básicos: interpretación, análisis, evaluación y autorregulación. La clara semejanza entre estas dos taxonomías integra la postura pluralista de la investigación del PC y avanza la sistematización del constructo PC. Las anteriores dimensiones y categorías del PC no deben interpretarse como elementos aislados o separados; por el contrario: todas las dimensiones pueden estar relacionadas, interactuar y aportar entre sí, pues los procesos de pensamiento son cogniciones que integran una variedad de destrezas combinadas.

El aspecto educativo más importante del PC es la relación positiva entre el pensamiento y el aprendizaje, un lugar común de la psicología cognitiva, ya que ambos implican múltiples componentes y niveles, cognitivos y no cognitivos, interdependientes e interrelacionados. Además, diversas evidencias empíricas confirman la dinámica interactiva entre el PC y otras variables educativas, que contribuyen a desarrollar mejor el aprendizaje, especialmente los aprendizajes científicos STEM (Ford y Yore, 2014; Hattie, 2009; 2012; Phan, 2010; Torres y Solbes, 2016).

En suma, la escasa atención a la evaluación del PC en los niveles educativos tempranos con estudiantes más jóvenes sugiere desarrollar esta línea mediante una prueba de evaluación del PC apropiada, no solo centrada en la personalidad de los jóvenes, sino también en destrezas específicas de interés para la práctica educativa en esos niveles tempranos, especialmente en áreas científicas (STEM) por la identidad de las destrezas evaluadas con el pensamiento científico (Manassero y Vázquez, 2020a).

El objetivo de investigación de este estudio es diagnosticar el PC en estudiantes de los

años finales de la Educación Secundaria Obligatoria, para lo cual desarrolla un instrumento de evaluación apropiado, investiga su relación con el aprendizaje (las calificaciones escolares) y compara los niveles de las destrezas de PC, como criterios externos de validez empírica; además, se presentan métodos analíticos confirmatorios acerca de la validez y la fiabilidad psicométricas del instrumento.

Materiales y métodos

Participantes

Los participantes son seis grupos naturales de estudiantes pertenecientes a tres centros educativos, uno privado concertado y dos públicos, situados en ciudades pequeñas. Fueron seleccionados por el interés del profesorado en el diagnóstico de las destrezas de PC, que dirigieron la aplicación de las pruebas. La muestra válida está formada por 88 estudiantes de los cursos tercero y cuarto de la Educación Secundaria Obligatoria (en adelante, ESO) del sistema educativo español, 48 hombres y 40 mujeres con edades comprendidas entre 14 y 17 años (promedio 15,4 años). Los estudiantes no habían recibido instrucción previa sobre destrezas de pensamiento, de manera que la prueba se planteó como un diagnóstico inicial.

Instrumento

El instrumento actual, Retos de pensamiento (RdP) es el resultado de varios procesos de desarrollo previos: diseño de un conjunto amplio de preguntas, elaboración de un banco de ítems de PC, aplicación y revisión piloto de resultados psicométricos y creación del instrumento actual, por selección y adaptación al nivel de los participantes (Manassero y Vázquez, 2020b; 2020c).

Las destrezas de PC evaluadas con el RdP se decidieron desde la práctica educativa, por el interés específico de un centro participante: deducción (dimensión razonamiento), asunciones (dimensión evaluación), secuenciación (dimensión creatividad) y toma de decisiones (dimensión procesos complejos). El contenido de las preguntas se diseñó atendiendo a criterios habituales en construcción de tests, generales (relevancia, representatividad, diversidad, inteligibilidad, claridad y sencillez) y específicos (la demanda cognitiva va dirigida a la destreza que representa, plantea un reto motivador y está evolutivamente adaptada al nivel de los participantes). La demanda cognitiva requerida para dar la respuesta correcta es un auténtico desafío independiente de la cultura, lo cual significa que la prueba presenta contenidos comunes a todas las personas, de modo que la elaboración de la respuesta correcta no es influida por conocimientos curriculares o culturales previos, y, por tanto, la prueba es aceptable para jóvenes de diferentes culturas y las medidas resultantes no están sesgadas por la influencia de una cultura o unos conocimientos específicos.

El instrumento RdP está formado por 23 ítems orientados a valorar las destrezas planteadas, combinando cuestiones verbales y figurativas, aquellas extraídas de test estandarizados y estas de elaboración propia (tabla 1). Las preguntas de las destrezas de deducción y asunciones se han seleccionado del test de Cornell, una narración ficticia sobre unos exploradores que llegan al planeta Nicoma, cuyas preguntas se responden con base en la información desplegada en la historia (Ennis y Millman, 2005). Un reactivo verbal sobre toma de decisiones y otras nueve preguntas figurativas de elaboración propia plantean contenidos para las destrezas de secuenciación y decisiones, que hacen estas preguntas más independientes de la cultura y los conocimientos escolares, y promueven mejor la motivación, la comprensión, el sentido de reto y la agilidad de las respuestas (véase el test RdP en el anexo).

Los formatos de respuesta son cerrados por sus ventajas: puntuaciones estandarizadas y objetivadas, elaboración y análisis de los resultados y líneas base para comparaciones rápidas, evaluación de cada destreza válida y fiable, y mejor ajuste de la demanda cognitiva y transferencia a la práctica educativa para uso por profesores.

Tabla 1. *Tabla de especificaciones de las cuatro destrezas de pensamiento crítico evaluadas a partir del instrumento RdP*

Destrezas de pensamiento	Fuente	Tipo	Ítems	Rango	Mínimo-máximo	Media	Desv. est.	Fiabilidad[a]
Deducción	Cornell (Nicoma)	Verbal	6	0-6	1-6	3,52	1,11	,114
Asunciones	Cornell (Nicoma)	Verbal	5	0-5	0-5	2,10	1,24	,432
Secuenciación	Elaboración propia	Figurativo	6	0-6	0-6	3,68	1,59	,800
Decisiones	Elaboración propia	Situaciones	6	0-6	0-4	2,17	1,35	,952
Global			23	0-23	5-19	11,48	3,61	,805

[a] Coeficientes de fiabilidad *EAP (expected a posteriori)* calculados con un método robusto (RULS) basado en correlaciones policóricas para datos bimodales.

Procedimiento

El instrumento RdP fue aplicado sin límite de tiempo a los estudiantes por su profesor, en su grupo de clase al final de curso, utilizando dispositivos digitales y siguiendo directrices comunes; para incentivar el esfuerzo y la motivación se planteó como una prueba de evaluación. Las respuestas correctas reciben un punto, y las respuestas incorrectas cero, 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 conforman y la puntuación global de pensamiento es la suma de los aciertos totales (tabla 1), que puede considerarse una estimación del nivel de PC de los estudiantes, con base en las cuatro destrezas.

La validez de contenido se basa, por un lado, en la validez de las fuentes de procedencia: las pruebas estandarizadas en las preguntas prestadas y las publicaciones especializadas en las de elaboración propia; por otro, los criterios de mejor ajuste ítem-destreza y demanda cognitiva-nivel educativo de los estudiantes se han basado en el juicio profesional y el acuerdo entre investigadores para la selección

de preguntas. La validez convergente y discriminante del RdP se evalúa con las calificaciones escolares (criterio externo de validez) y la discriminación entre grupos.

La base de datos se ha sometido a control de calidad y procesado previos con SPSS (versión 25). El programa Factor aplica un método robusto de mínimos cuadrados no ponderados (RULS) a correlaciones policóricas, apropiadas para los datos bimodales (0-1) de las respuestas, y para la validez basada en los análisis factoriales exploratorios (AFE) y confirmatorios. La fiabilidad se calcula mediante el estadístico esperado *a posteriori* (EAP), mediante AFE, Unrestricted Factor Analysis y RULS (Ferrando y Lorenzo-Seva, 2017, 2018).

Resultados y análisis

Los descriptores estadísticos de las puntuaciones de las 23 variables obtenidos a partir de las respuestas de los estudiantes a la prueba RdP se resumen en la tabla 2. Las tasas de aciertos muestran una distribución equilibrada entre preguntas fáciles y difíciles, pues la mayoría de las preguntas (16) logran índices de facilidad intermedios (0,30 a 0,70), en tanto que una minoría (4) son muy fáciles (> 0,70) y otra minoría (3) muy difíciles (< 0,30). El promedio de aciertos global logra un valor adecuado próximo al 50 % (0,499), lo que confirma la dificultad media del instrumento RdP.

Tabla 2. *Proporción de aciertos en cada una de las 23 variables evaluadas con el instrumento RdP*

Ítems/Variables	Aciertos (0-1)	Desviación estándar
DEDUC1	0,8864	0,31919
DEDUC2	0,5682	0,49817
DEDUC3	0,4205	0,49646
DEDUC4	0,5000	0,50287
DEDUC5	0,2727	0,44791
DEDUC6	0,8750	0,33261
ASUNC1	0,3409	0,47673
ASUNC2	0,3977	0,49223
ASUNC3	0,5455	0,50078
ASUNC4	0,3636	0,48380
ASUNC5	0,4545	0,50078
SECUE1	0,9318	0,25350
SECUE2	0,5568	0,49961
SECUE3	0,7273	0,44791
SECUE4	0,5795	0,49646
SECUE5	0,4205	0,49646
SECUE6	0,4659	0,50170
DECIS1	0,4205	0,49646
DECIS2	0,5568	0,49961
DECIS3	0,4886	0,50274
DECIS4	0,3523	0,48042
DECIS5	0,1705	0,37819
DECIS6	0,1818	0,38790

Fuente: elaboración propia.

Las puntuaciones medias de las destrezas también están próximas al punto medio del rango total de cada una; por debajo del punto medio en las destrezas Asunciones y Toma de decisiones, y ligeramente por encima en Deducción y Secuenciación.

Tabla 3. *Estadística básica descriptiva de las calificaciones obtenidas en las asignaturas escolares por los estudiantes participantes en este estudio (n = 88)*

Asignaturas curriculares	Mínimo	Máximo	Media	Desviación estándar
Biología y Geología	3	10	7,01	1,568
Educación Física	3	10	7,38	1,549
Física y Química	3	10	7,53	1,735
Geografía e Historia	5	10	7,19	1,388
Lengua Castellana	3	10	6,70	1,669
Lengua Catalana	1	10	6,34	1,929
Matemáticas	2	10	6,59	1,779
Religión-Valores	2	10	7,13	1,622
Lengua Extranjera	2	10	7,02	1,849
Nota media			6,955	1,312

Fuente: elaboración propia.

Los parámetros estadísticos de las calificaciones escolares (rango 1-10) obtenidas por los estudiantes en sus asignaturas muestran cierta asimetría en las puntuaciones máximas y mínimas, pues mientras las primeras se alcanzan en todas las asignaturas, las segundas se distribuyen irregularmente. La asignatura Física y química tiene la puntuación media más alta de la muestra, mientras que la más baja corresponde a lengua catalana. Estos resultados están resumidos en la tabla 3.

Validez de criterio

Las correlaciones empíricas de las variables de pensamiento crítico entre sí y con las calificaciones de los estudiantes son la base de la validez convergente.

Correlaciones entre las destrezas de pensamiento

La literatura atribuye al dominio de destrezas de PC un impacto transversal en el aprendizaje, de modo que sus correlaciones positivas entre sí y, especialmente, con variables de aprendizaje (calificaciones escolares) verifican su validez convergente.

Todas las destrezas se correlacionan positivamente entre sí; además, la destreza Secuenciación (figurativa) presenta correlaciones estadísticamente significativas con las otras tres, siendo la más alta con Toma de decisiones, mientras que Asunciones no se correlaciona significativamente con las otras dos (tabla 4).

Tabla 4. *Correlaciones de Pearson entre las destrezas de pensamiento crítico (deducción, asunciones, secuenciación y toma de decisiones)*

Destrezas	Asunciones		Secuencia		Decisiones	
	Correlación	(Sig.)	Correlación	(Sig.)	Correlación	(Sig.)
Deducción	0,160	(0,136)	0,225*	(0,035)	0,269*	(0,011)
Asunciones			0,361**	(0,001)	0,134	(0,215)
Secuencia					0,460**	(0,000)

** La correlación es significativa en el nivel 0,05 (bilateral).*

*** La correlación es significativa en el nivel 0,01 (bilateral).*

Fuente: elaboración propia.

Correlaciones entre destrezas de pensamiento y calificaciones escolares

Las correlaciones de Pearson positivas entre variables de PC (destrezas) y de aprendizaje (calificaciones escolares), como criterios externos y diferenciados del PC, verifican la validez predictiva de la prueba.

Tabla 5. *Correlaciones de Pearson entre las puntuaciones de las cuatro destrezas de pensamiento crítico, la puntuación global de pensamiento, las calificaciones, las asignaturas escolares y la nota media de estas calificaciones escolares*

Calificaciones	Pensamiento				
	Global	Deducción	Asunciones	Secuencia	Decisiones
Biología y Geología	0,250*	0,226	0,096	0,078	0,320**
Educación Física	0,241*	0,205	0,004	0,007	0,464**
Física y Química	0,258*	0,098	0,140	0,083	0,383**
Geografía e Historia	0,232*	0,068	0,128	0,112	0,313**
Lengua Castellana	0,241*	0,232*	0,120	0,025	0,314**
Lengua Catalana	0,331**	0,269*	0,163	0,148	0,339**
Matemáticas	0,321**	0,196	0,149	0,198	0,326**
Religión-Valores	0,200	0,087	0,109	0,094	0,252*
Lengua Extranjera	0,372**	0,225*	0,135	0,247*	0,393**
Nota media	0,350**	0,225*	0,162	0,144	0,430**

** La correlación es significativa en el nivel 0,05 (bilateral).*

*** La correlación es significativa en el nivel 0,01 (bilateral).*

Fuente: elaboración propia.

Todas las correlaciones entre destrezas del PC y calificaciones son positivas y la mayoría significativas, lo cual verificaría la validez de constructo, aunque el análisis de algunos matices es interesante. Las dos variables globalizadoras (puntuación global de PC y nota media de calificaciones) exhiben correlaciones altas y significativas con las demás variables; las más altas de la puntuación global del PC se dan con las calificaciones de Lengua extranjera, Lengua catalana y

Matemáticas; la nota media de calificaciones muestra el índice de correlación más alto con Toma de decisiones y la puntuación global del PC, mientras no son significativas con Asunciones y Secuenciación (tabla 5).

Las correlaciones de destrezas y calificaciones muestran algunos patrones específicos. Toma de decisiones se correlaciona significativamente con todas las asignaturas y Deducción con las tres asignaturas de lenguas; las correlaciones de Asunciones y Secuenciación no son significativas con casi todas las asignaturas.

En resumen, las correlaciones positivas, y en muchos casos significativas, entre destrezas y calificaciones de aprendizajes escolares verifican la validez convergente de la prueba, aunque también sugieren la importancia discriminante de las destrezas del PC en relación con los diversos aprendizajes escolares.

Análisis comparativo de las destrezas de pensamiento entre dos grados escolares

Aunque los participantes no reciben enseñanza específica sobre pensamiento, el mero desarrollo evolutivo y los aprendizajes a lo largo de un curso adicional hacen plausible hipotetizar que los participantes del curso superior (cuarto, ESO4) logren puntuaciones de PC mayores que los participantes del grado inferior (tercero, ESO3) como prueba de la capacidad discriminante de la prueba.

La figura 7 presenta las puntuaciones medias de las destrezas de pensamiento entre los dos grupos comparados; los estudiantes de cuarto de ESO obtienen puntuaciones mayores que los de tercero, con excepción de la destreza Asunciones, aunque las diferencias no son estadísticamente significativas en ninguna de las cinco variables, contrastadas con una prueba Anova para comparar ambos grupos.

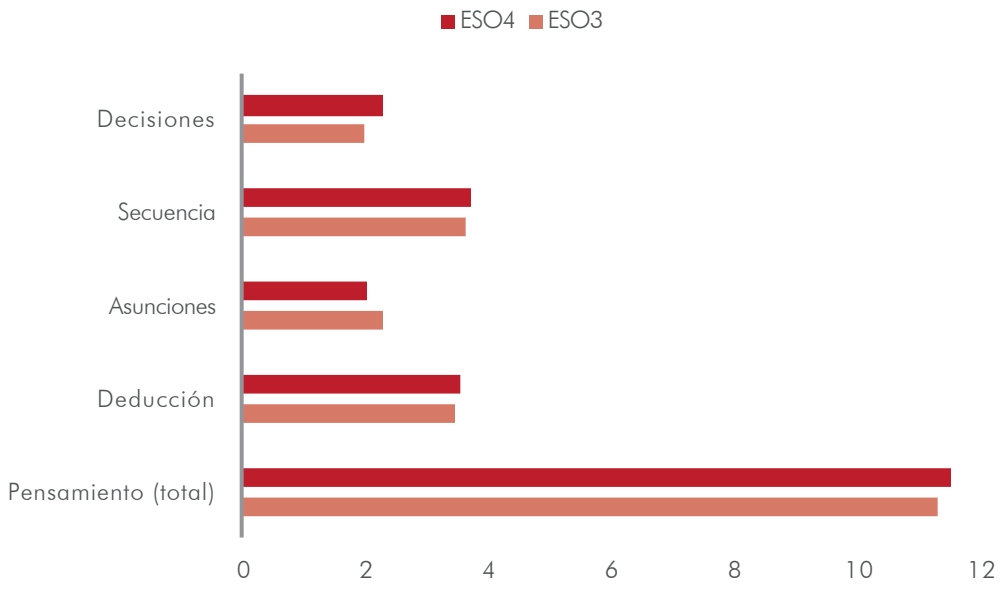


Figura 1. *Puntuaciones medias de las destrezas de pensamiento para los dos grupos de participantes en los cursos tercero (ESO3, n = 28) y cuarto (ESO4, n = 60) de ESO.*

Fuente: elaboración propia.

Este análisis comparativo demuestra sensibilidad del instrumento de PC para detectar diferencias evolutivas entre los dos grupos comparados y con el sentido y magnitud esperados: el instrumento asigna puntuaciones mayores a los estudiantes del grado superior (cuarto de ESO) y no produce diferencias significativas entre grados (pues los participantes no recibieron instrucción sobre destrezas de pensamiento). Ambos resultados apoyan la validez discriminante del instrumento, aunque la inversión de las diferencias en la destreza asunciones aporta un indicador adicional para su revisión.

Análisis de regresión entre pensamiento y calificaciones

Con el índice de correlaciones se puede conocer la varianza común entre una pareja de variables, pero no permite contrastar la potencia predictiva de una variable en comparación con otras. Para comparar esta potencia se ha realizado un análisis de regresión lineal entre las destrezas de pensamiento y las calificaciones escolares, alternando esos dos conjuntos de variables como variables dependientes o como predictores (tablas 6 y 7).

Tabla 6. *Coeficientes estandarizados (beta) de los predictores (calificaciones de las asignaturas) en el análisis de regresión estadística lineal respecto a las variables dependientes (destrezas de pensamiento)*

Variables	Predictores (destrezas de pensamiento)					Varianza
Dependientes (asignaturas)	Deducción	Asunciones	Secuencia	Decisiones	Total	común (%)
Biología y Geología			-	0,365**		140,6
Educación Física			-0,282*			280,2
Física y Química	-		-	0,444***		170,5
Geografía e Historia	-		-	0,342**		110,1
Lengua Castellana			-	0,354**		150,9
Lengua Catalana			-			160,3
Matemáticas				0,281*		120,8
Religión-Valores			-			70,3
Lengua Extranjera				0,333**		170,6
Nota media				0,443***		210,9

* Coeficiente estandarizado significativo al nivel 0,05

** *Ídem al nivel 0,01*

*** *Idem al nivel 0,001*

(-) Coeficiente estandarizado negativo (no significativo)

Fuente: elaboración propia.

La tabla 6 muestra que la asignatura Educación Física alcanza un alto valor de la varianza común con las cuatro destrezas (28,2 %) y se alcanzan valores importantes en la mayoría de las asignaturas (11 a 22 %). La destreza Toma de decisiones es el predictor más importante de las calificaciones, pues resulta positivo y significativo en la mayoría de las asignaturas (7 de 9), mientras que el

resto de las destrezas no alcanzan significación (excepto Secuencia para Educación Física). Otro hecho para destacar es el signo negativo de la destreza Secuencia como predictor no significativo de la mayoría de las asignaturas, excepto dos (Matemáticas y Lengua Extranjera), lo que podría interpretarse como ausencia de esa destreza en la Educación.

Tabla 7. *Coeficientes estandarizados (beta) de los predictores (calificaciones de las asignaturas) en el análisis de regresión estadística lineal respecto a las variables dependientes (destrezas de pensamiento)*

Predictores (asignaturas)	Variables dependientes (destrezas de pensamiento)				
	Deducción	Asunciones	Secuencia	Decisiones	Global
Biología y Geología		-	-		
Educación Física		-	-	0,396*	-
Física y Química				0,393*	
Geografía e Historia	-0,580*			-	-
Lengua Castellana		-	-	-	-
Lengua Catalana				-	
Matemáticas				-	
Religión-Valores	-		-		-
Lengua Extranjera			0,515*		0,415*
Varianza común (%)	19,4	6,1	21,2	31,8	19,5

** Coeficiente estandarizado significativo al nivel ,05*

*** Idem al nivel 0,01*

**** Idem al nivel 0,001*

(-) Coeficiente estandarizado negativo (no significativo)

Fuente: elaboración propia.

La predicción de PC por las calificaciones muestra que Toma de decisiones presenta un alto valor de la varianza común por las calificaciones (31,8 %), mientras que Asunciones obtiene el mínimo (6,1 %), alcanzando las otras destrezas valores relevantes de varianza común (en torno al 20 %). Sin embargo, pocas asignaturas son predictores significativos de las destrezas; Física y Química y Educación Física son predictores positivos y significativos de la Toma de decisiones, mientras que Lengua Extranjera predice significativamente la destreza Secuencia y la puntuación global del PC; además, solo Física y Química y Lengua Extranjera exhiben coeficientes positivos en todos los casos. La asignatura geografía e historia se destaca por su alto coeficiente significativo, pero negativo, como predictor de Deducción, lo que indicaría que los estudiantes con altas calificaciones en esa asignatura tienen bajas puntuaciones en Deducción. Estos resultados sugieren una contribución diferencial de las distintas asignaturas a las destrezas.

Fiabilidad del instrumento de evaluación de pensamiento crítico

La fiabilidad del instrumento RdP se calcula con el coeficiente EAP (*expected a posteriori*), calculado con un método RULS y correlaciones policóricas para datos bimodales (tabla 1). El coeficiente para la escala global de PC es bueno; sin embargo, la fiabilidad de las escalas que representan las cuatro destrezas de PC es muy

diferente. Las destrezas de Deducción y Asunciones presentan coeficientes bajos, mientras que los coeficientes de Secuenciación y Toma de decisiones son excelentes.

Aunque se conoce el descenso automático de la fiabilidad por bajo número de ítems, los resultados indican la necesidad de revisar estas escalas para mejorar su fiabilidad, aumentando tanto la calidad como la cantidad de ítems. Por ejemplo, suprimiendo tres ítems en Deducción y Asunciones se obtienen ya valores de fiabilidad moderados (0,602 y 0,617) con los mismos datos, a pesar de disminuir drásticamente su número.

Validez de constructo: análisis factorial exploratorio y confirmatorio

Se aplica el método robusto RULS para verificar la validez del instrumento contrastando, en primer lugar, el grado en que cada destreza constituye un constructo unidimensional, y, en segundo lugar, la bondad del ajuste de todas las preguntas para representar las cuatro destrezas definidas teóricamente.

Para verificar la hipótesis unidimensional en cada una de las destrezas, las preguntas de cada destreza se han sometido por separado a un AFE de factor único (tabla 8).

Tabla 8. *Análisis factorial exploratorio de componentes principales de las preguntas correspondientes a cada una de las destrezas sometidas a un modelo de factor único*

Estadísticos	Deducción		Asunciones		Secuencia		Decisiones	
	Variables	Cargas	Variables	Cargas	Variables	Cargas	Variables	Cargas
	DEDUC1	1,000	ASUNC1	.346	SECUE1	.711	DECIS1	.968
	DEDUC2	-0,025	ASUNC2	0,104	SECUE2	0,373	DECIS2	0,876
	DEDUC3	0,088	ASUNC3	0,417	SECUE3	0,630	DECIS3	0,754
	DEDUC4	0,389	ASUNC4	0,477	SECUE4	0,673	DECIS4	-0,140
	DEDUC5	-0,020	ASUNC5	0,310	SECUE5	0,706	DECIS5	0,187
	DEDUC6	0,407			SECUE6	0,502	DECIS6	0,490
KMO	0,177		0,514		0,735		0,627	
Bartlett (p)	0,000		0,023		0,000		0,000	
% varianza	31,2		29,4		47,1		47,8	

KMO: Prueba Kaiser-Meyer-Olkin.

Bartlett: Prueba de esfericidad de Bartlett.

Fuente: elaboración propia.

La prueba de esfericidad de Bartlett es significativa en todas las destrezas y los valores de Kaiser-Meyer-Olkin son aceptables en dos destrezas (Secuencia y Toma de decisiones), moderado en asunciones y bajo en Deducción. El porcentaje de varianza común es alto para Secuencia y Toma de decisiones y más bajo en Asunciones y Deducción.

El análisis de las cargas de cada una de las variables sobre el factor único de cada destreza permite identificar cuestiones deficientes según las aportaciones a la varianza explicada. En la destreza figurativa de Secuenciación todas las preguntas aportan al factor único con cargas apreciables. Análogamente, cuatro preguntas de Asunciones presentan también cargas altas y positivas (una más baja), Deducción tiene tres preguntas con cargas deficientes y Toma de decisiones muestra dos preguntas con cargas bajas. Todas estas preguntas deberían revisarse para mejorar el instrumento.

En segundo lugar, con todas las preguntas del instrumento se ha ensayado por el método RULS un AFE de cuatro factores, para confirmar la estructura teórica de cuatro destrezas. Los parámetros de la matriz de correlaciones policóricas son adecuados (KMO = 0,703; Bartlett, $p < 0,000$), y muestra cuatro autovalores mayores que la unidad y una buena proporción de la varianza explicada (46 %) por cuatro factores empíricos de este modelo. Una rotación Promin robusta para lograr la simplicidad de los factores muestra la matriz de cargas factoriales de la tabla 9 (Lorenzo-Seva y Ferrando, 2019).

Tabla 9. Matriz de cargas de los factores rotados resultantes de un AFE y rotación Promin con método robusto de mínimos cuadrados no ponderados (RULS) para un modelo de cuatro factores realizado sobre todas las cuestiones

Variables	Factor 1	Factor 2	Factor 3	Factor 4	Autovalores
DEDUC1		0,446	0,352		3,591
deduc2			0,464	0,385	2,132
DEDUC3			0,337		1,439
DEDUC4			0,321		1,188
DEDUC5	0,691		-0,367		0,838
DEDUC6		0,780		-0,458	0,586
ASUNC1				-0,316	0,533
ASUNC2			0,760		0,457
ASUNC3					0,279
ASUNC4	0,350				0,247
ASUNC5	0,482				0,146
SECUE1		0,583			0,103
SECUE2	0,538				0,058
SECUE3	0,442				-0,051
secue4		0,317			-0,117
SECUE5		0,635			-0,143
SECUE6		0,444			-0,211
DECIS1				0,747	-0,274
DECIS2				0,532	-0,384
DECIS3				0,505	-0,417
DECIS4					-0,519
DECIS5	0,654	0,301			-0,532
DECIS6				-0,711	-0,601
Fiabilidad ORION[a]	,805	,853	,765	,846	
% varianza acumulada	18,0	29,2	38,3	46,0	

Cargas < 0,30 borradas

a Overall Reliability of fully-Informative prior Oblique N-EAP scores.

Fuente: elaboración propia.

Los resultados empíricos del AFE ajustan el número teórico de factores (4) y la matriz de cargas indica que los factores aproximan razonablemente el modelo teórico, pues las preguntas con mayores cargas sobre los cuatro factores empíricos se corresponden con las destrezas teóricas. El factor 1 correspondería a la destreza Asunciones, aunque ASUNC2 (ya identificado por su mal funcionamiento) presenta una carga cruzada sobre el factor 3. El factor 2 corresponde a Secuencia, aunque SECUE2 y SECUE3 presentan cargas cruzadas sobre el factor 1. El factor 3 corresponde a Deducción, aunque DEDUC5 presenta carga negativa y DEDUC6 carga cruzada sobre el factor 2. El factor 4 corresponde a la destreza Decisiones, aunque DECIS6 presenta carga negativa. Otras preguntas presentan cargas negativas o cruzadas entre factores disfuncionales, lo que sugiere también una revisión para mejora.

En resumen, el conjunto de preguntas asignadas teóricamente a cada destreza presenta estructura unidimensional, lo cual apoyaría la validez de cada una de las escalas, como constructos con identidad propia para representar cada destreza, aunque las preguntas con cargas bajas sobre el factor único deben revisarse. La estructura empírica del AFE completo ajusta razonablemente el modelo teórico de cuatro destrezas, pues los cuatro factores empíricos obtenidos reproducen las destrezas teóricas, aunque cada factor aparece contaminado con cargas cruzadas hacia otras destrezas, lo que sugiere la revisión de varias preguntas (DEDUC2, DEDUC3 DEDUC5, ASUNC2, DECIS4, DECIS5 Y DECIS6).

Estadística confirmatoria de los modelos de factores

En la tabla 10 se resumen los parámetros de bondad de ajuste robusta de la estadística confirmatoria para los modelos de factores unidimensionales, en cada una de las destrezas, y del modelo de cuatro factores, para el instrumento total.

El modelo empírico de cuatro factores para el instrumento total presenta parámetros confirmatorios favorables que corroboran su validez y fiabilidad como instrumento de evaluación global del PC (tabla 9). Análogamente, los cuatro modelos unidimensionales de cada una de las destrezas teóricas tienen buenos índices de bondad de ajuste (GFI), aunque los restantes parámetros de ajuste son más variables. La destreza teórica Decisiones es la que presenta un mejor conjunto global de parámetros de bondad de ajuste y la destreza Deducción, el más pobre.

Tabla 10. *Parámetros estadísticos de la bondad de ajuste robusta confirmatoria de los modelos de factores*

Destrezas	Modelos	Medidas de ajuste absoluto			Medidas de ajuste incremental		
	Nº Ítems (Factores)	Ji-cuadrado	p	RMSEA[A]	CFI[B]	NFI[C]	GFI[D]
Deducción	6(F3)	43,032	0,000	0,179	0,611	0,352	0,856
Suposiciones	5(F1)	8,623	0,128	0,088	0,697	0,393	0,964
Secuencias	6(F2)	47,666	0,000	0,202	0,848	0,746	0,902
Decisiones	6(F4)	28,175	0,001	0,139	0,931	0,885	0,943
Total	23(4)	135,033	0,967	0,021	0,990	0,985	0,928

[a] *Root Mean Square Error of Approximation*
[b] *Comparative Fit Index*
[c] *Normed Fit Index*
[d] *Goodness of Fit Index*
Fuente: elaboración propia.

Por otro lado, los parámetros sensibles a la muestra de la estadística confirmatoria (ji-cuadrado y CFI) no empeoran los parámetros insensibles a la muestra (RMSEA, CFI, GFI), de modo que la muestra pequeña usada no parece haber influido en los resultados obtenidos, aunque aumentar el tamaño de la muestra es otro elemento de mejora obvio.

Discusión y conclusiones

El objetivo de este estudio es desarrollar un instrumento (RdP) para evaluar cuatro destrezas del PC (Deducción, Asunciones, Secuenciación y Toma de decisiones) solicitadas por los educadores, adaptadas a las necesidades de estudiantes jóvenes (de quince y dieciséis años), libres de cultura que intentan superar limitaciones de los instrumentos estandarizados de PC (uso en adultos). Para ello, las preguntas de las dos primeras destrezas han sido elegidas de pruebas estandarizadas (Ennis y Millman, 2005) y otras, para evaluar las dos últimas destrezas, han sido adaptadas por los autores. El principal hallazgo es que la validez y fiabilidad de la prueba global para medir PC son aceptables, pues su puntuación se correlaciona positiva y significativamente con las calificaciones escolares, alcanza un valor excelente de fiabilidad y los parámetros confirmatorios de bondad de ajuste son buenos. La correlación positiva entre puntuaciones del instrumento y calificaciones escolares verifica empíricamente el carácter transversal del PC respecto a los distintos aprendizajes escolares, sostenido por algunos agentes (European Union, 2014), aunque esta interpretación requiere nuevos estudios con pruebas mejoradas y muestras amplias.

La validez de las escalas que representan las cuatro destrezas teóricas es aceptable, porque cada destreza es unidimensional y las correlaciones con las calificaciones, aunque positivas solo son significativas en Toma de decisiones. Sin embargo, la estadística confirmatoria de la bondad de ajuste es más heterogénea, pues el AFE exhibe cargas cruzadas entre factores y el AFE unidimensional presenta bajas cargas en algunas cuestiones. La fiabilidad de las destrezas, aunque minorada por el automático impacto a la baja de su reducida longitud (5-6 ítems), es buena para las escalas de Secuenciación y Decisiones. Este estudio innova el diseño de instrumentos para evaluar PC atendiendo a las necesidades de la práctica escolar en la educación obligatoria,

adaptados a los perfiles, las necesidades y la edad de los estudiantes y libres de cultura, pues las medidas son independientes de la cultura o los conocimientos previos de los encuestados. También es destacable que los reactivos figurativos (Secuencia) muestran mejor fiabilidad y validez que los verbales, apuntando una línea innovadora para el desarrollo de pruebas de PC, especialmente para edades donde los componentes verbales y los conocimientos específicos pueden condicionar decisivamente la elaboración de respuestas. La valoración de PC con reactivos figurativos y libres de cultura sigue la tendencia actual de evitar la interferencia del conocimiento previo para facilitar su transferencia a otros contextos, pero su mejora requiere desarrollar nuevas cuestiones y aplicaciones en esta línea (OECD, 2018; Unesco, 2015).

Las principales limitaciones que plantean los resultados es la revisión de las preguntas disfuncionales, aumentar su número, y ampliar la muestra para aplicar modelos de ecuaciones estructurales apropiado para las variables dicotómicas empleadas (Muñiz y Fonseca-Pedrero, 2019). El análisis de regresión muestra un resultado positivo y singular: la gran proporción de varianza común entre los aprendizajes escolares y las destrezas de pensamiento. Al tiempo, muestra una débil aportación de las diferentes asignaturas a las destrezas de pensamiento, y, en sentido inverso, solo la destreza Toma de decisiones predice significativamente la calificación en la mayoría de las asignaturas. En consecuencia, las aportaciones no significativas o negativas sugieren carencias de destrezas en algunas asignaturas que requieren un análisis más profundo.

Finalmente, Scriven y Blair (2021) recomiendan al profesorado aplicar los marcos mentales elaborados por Hattie (2009), el primero de los cuales es creer firmemente que su tarea fundamental es evaluar el efecto de su enseñanza en el aprendizaje de los estudiantes, un consejo extremadamente desafiante si se tiene en cuenta la carencia de test de evaluación de PC para la población escolar más joven.

Este estudio trata de ser lo suficientemente simple para que el instrumento RdP pueda ser utilizado por profesores y estudiantes en las aulas, ofrecer un lenguaje comprensible para profesores y alumnos, reflejar un marco conceptual comprensible y consensuado internacionalmente, ayudar a los profesores a diseñar actividades y ejercicios pedagógicos relacionados con el currículo que imparten y con demostrar el logro de las destrezas y dar una descripción positiva del progreso en las destrezas seleccionadas y, a los estudiantes, un marco de referencia para su propia autoevaluación. La recomendación anterior también sugiere la necesidad de promover investigaciones en las que se elaboren instrumentos de evaluación de PC para escolares, sustentada en la importancia transversal y transferible del PC para la educación general y para las áreas STEM, cuyos aprendizajes parecen especialmente más empapados de destrezas de PC (Erduran y Kaya, 2018; Tamayo, 2017; Tenreiro-Vieira y Vieira, 2014; Torres y Solbes, 2016).

Referencias

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 XXI*, *23*(1), 45-74. 10.5944/educXX1.23853.

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.

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 Millman, J. (2005). *Cornell Critical Thinking Test, Level X*. The Critical Thinking Company.

Erduran, S. y Kaya, E. (2018). Drawing nature of science in pre-service science teacher education: Epistemic insight through visual representations. *Research in Science Education*, *48*(6), 1133-1149. https://doi.org/10.1007/s11165-018-9773-0.

European Union. (2014). *Key competence development in school education in Europe. KeyCoNet's review of the literature: A summary*. 2014. http://keyconet.eun.org.

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. 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. doi: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* 2.ª ed. (pp. 7-26). University of Windsor.

Ford C. L. y Yore, L. D. (2014). Toward convergence of critical thinking, metacognition, and reflection: Illustrations from natural and social sciences, teacher education, and classroom practice. En A. Zohar y Y. J. Dori (Eds.), *Metacognition in Science Education* (pp. 251-271). Springer.

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

Halpern, D. F. (2010). *Halpern critical thinking assessment*. Schuhfried.

Hattie, J. (2009). *Visible learning: A synthesis of over 800 meta-analyses relating to achievement*. Routledge.

Hattie, J. (2012). *Visible learning for teachers: Maximizing impact on learning*. Routledge.

International Society for Technology Education. (2003). *National Educational Technology Standards for Teachers: Preparing teachers to use technology*. International Society for Technology Education.

Krathwohl, D. (2002). A revision of Bloom's taxonomy. *Theory into Practice*, *41*, 212-218.

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. doi: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 (Org.), *Ciência, Tecnologia & Sociedade: Ensino, Pesquisa e Formação* (pp. 17-38). Unicsul.

Manassero-Mas, M. A. y Vázquez-Alonso, Á. (2020a). Pensamiento científico y pensamiento crítico: Competencias transversales para aprender. *Indagatio,* 12, 401-419. https://doi.org/10.34624%2Fid.v12i4.21808.

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

Manassero-Mas, M. A. y Vázquez-Alonso, A. (2020c). Las destrezas de pensamiento y las calificaciones escolares en educación secundaria: Validación de un instrumento de evaluación libre de cultura. *Tecné, Epistemé y Didaxis: TED, 48*. https://doi.org/10.17227/ted.num47-9801.

McDonald, C. V. y McRobbie, C. J. (2012). Utilising argumentation to teach nature of science. En B. J. Fraser, K. G. Tobin y C. J. McRobbie (Eds.), *Second international handbook of Science Education* (pp. 969-986). Springer.

Muñiz, J. y Fonseca-Pedrero, E. (2019). Diez pasos para la construcción de un test. *Psicothema*, *31*(1), 7-16.

National Research Council (2012). *Education for life and work: Developing transferable knowledge and skills in the 21st century*. The National Academies Press.

Organisation for Economic Co-operation and Development (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.

Phan, H. P. (2010). Critical thinking as a self-regulatory process component in teaching and learning. *Psicothema*, *22*, 284-292.

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

Rivas, S. F. y Saiz, C. (2012). Validación y propiedades psicométricas de la prueba de pensamiento crítico Pencrisal. *Revista Electrónica de Metodología Aplicada*, *17*, 18-34.

Saiz, C. (2017). *Pensamiento crítico y cambio*. Pirámide.

Scriven, M. y Blair, J. A. (2021). Teaching critical thinking. En J. A. Blair (Ed.), *Studies in critical thinking* (2.ª ed.) (pp. 31-35). University of Windsor.

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

Simonneaux, L. (2014). From promoting the techno-sciences to activism: A variety of objectives involved in the teaching of SSIs. En L. Bencze y S. Alsop (Eds.), *Activist Science and Technology Education* (pp. 99-112). Springer.

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

Tamayo, O. E. (2017). Interacciones entre naturaleza de la ciencia y pensamiento crítico en dominios específicos del conocimiento. *Enseñanza de las ciencias*, número extra (pp. 521-526).

Tenreiro-Vieira, C. y Vieira, R. M. (2014). *Construindo práticas didático-pedagógicas promotoras da literacia científica e do Pensamento Crítico*. Iberciencia.

Torres, N. Y. y Solbes, J. (2016). Contribuciones de una intervención didáctica usando cuestiones sociocientíficas para desarrollar el pensamiento crítico. *Enseñanza de las Ciencias*, *34*(2), 43-65.

Unesco (2015). Rethinking Education: Towards a global common good? http://unesdoc.unesco.org/images/0023/002326/232697s.pdf.

Valenzuela, J. (2008). Habilidades de pensamiento y aprendizaje profundo. *Revista Iberoamericana de Educación*, *46*.

Vázquez-Alonso, Á. y Manassero-Mas, M. A. (2018). Más allá de la comprensión científica: Educación científica para desarrollar el pensamiento. *Revista Electrónica de Enseñanza de las Ciencias, 17*(2), 309-336.

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

Walton, D. y Macagno, F. (2015). A classification system for argumentation schemes. *Argument and Computation*, *6*(3), 214-249.

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

Forma de citar este artículo

Manassero-Mas, M. A. y Vázquez-Alonso, A. (2023). Evaluación de destrezas de pensamiento en educación secundaria: propiedades psicométricas de un instrumento independiente de la cultura. *Tecné, Episteme y Didaxis: TED*, (53), 14-41. https://doi.org/10.17227/ted.num53-15830

Anexo

Retos de pensamiento

Esta prueba reta tus habilidades de pensamiento planteando problemas para resolver. Las respuestas no requieren conocimientos, solo requieren pensar. Es muy importante esforzarse en pensar intensamente para responder lo mejor posible.

El test es fruto del esfuerzo de muchas personas y sus resultados son valiosos para mejorar la educación. Muchos estudiantes han encontrado atractivos estos retos.

Debes responder individualmente y sin prisas. Emplea el tiempo que necesites para responder.

Sinceramente, gracias por tu esfuerzo y generosidad para responder los retos lo mejor posible.

Exploración en la aldea de Nicoma: ¿qué se puede hacer?

Estamos a mediados de junio del año 2050. Imagínese que pertenece al segundo grupo de habitantes de la Tierra trasladados al planeta Nicoma, recientemente descubierto, porque no se sabe nada sobre el primer grupo que aterrizó en Nicoma dos años antes. Su grupo fue enviado a hacer un informe sobre lo que sucedió al primer grupo.

Empieza a oscurecer, por lo tanto, acampan para pasar la noche. A la mañana siguiente, emprenden de nuevo su camino. Después de haber caminado durante una hora, el grupo llega a un pueblo de cabañas de piedra. La aldea parece vacía. El sol brilla intensamente. Como usted es el jefe del grupo, los otros miembros le traen informaciones.

Usted decide llevar a su grupo a la cima de la colina, que se encuentra detrás de la mayor de las cabañas, para ver si puede averiguar de dónde proviene el humo. En la distancia puede ver un grupo de unos 40 individuos reunidos alrededor de una fogata.

La gente alrededor de la hoguera se levanta y camina hacia la aldea. Rápidamente usted lleva a su pequeño grupo a otro lugar en la colina cercana. Desde allí se puede ver el pueblo sin ser visto. Quiere saber si la gente del pueblo no es hostil, si los exploradores están presos y cuántos de ellos quedan. El mecánico anota lo que las personas dicen ver.

Usted junto con su grupo va a intentar averiguar si los habitantes de la aldea son amistosos. Si fuesen hostiles, usted necesitará salvar a los exploradores. Trate de razonar acerca de las consecuencias.

En este folleto se tendrán en cuenta algunas de las cosas que su grupo descubrió en el planeta Nicoma. A continuación, se plantean preguntas que requieren un razonamiento claro. La historia tiene dos partes.

Responda las preguntas como si las cosas que se dicen o se informan en cada una siempre fueran ciertas.

Para cada tema de esta parte debe decidir la consecuencia verdadera que se sigue razonadamente de las declaraciones de la persona. Es decir, para cada pregunta suponga como cierto lo que la persona dice. Después, decida lo que es verdadero, como respuesta y consecuencia lógica de lo anterior.

Seleccione A, B o C (o deje en blanco si no sabe la respuesta). Considere cada pregunta independiente de las demás.

Nunca responda al azar. Si no sabe, deje en blanco la respuesta. Pero si tiene alguna buena idea, incluso sin estar seguro, responda a la pregunta según esa idea.

Aquí está un ejemplo:

51. El mecánico dice: "si estos seres son personas de la Tierra, también nos recibirán bien a nosotros. Son seguramente gente de la Tierra".

¿Cuál de las siguientes afirmaciones es cierta, como consecuencia de lo anterior?

A. Estos seres no nos reciben bien.

B. Estos seres no son de la Tierra.

C. Estos seres nos reciben bien.

Marque una respuesta.

La respuesta correcta es C. Si lo que dijo el mecánico es cierto, entonces también C debe ser cierto.

Para cada pregunta hay una respuesta que puede ser considerada *la consecuencia verdadera que se sigue de la afirmación de la persona*.

Considere solo una pregunta a la vez. En esta parte puede volver hacia atrás a una pregunta, bien sea para cambiarla o bien sea para dar una respuesta.

Ahora espere hasta que le ordenen empezar.

52. "Si estos seres son de la Tierra, entonces otra nave espacial más aterrizó en Nicoma.

Estos seres sin duda son gente de la Tierra".

¿Cuál de las siguientes afirmaciones es cierta, como consecuencia de lo anterior?

A. Otra nave aterrizó en Nicoma.

B. Estos seres no son de la Tierra.

C. No aterrizó otra nave espacial en Nicoma.

53. "Si estos seres son de la Tierra, entonces otra nave espacial más aterrizó en Nicoma. Pero ninguna otra nave aterrizó en Nicoma".

¿Cuál de las siguientes afirmaciones es cierta, como consecuencia de lo anterior?

A. Otra nave espacial aterrizó en Nicoma.

B. Estos seres no son de la Tierra.

C. Estos seres vinieron aquí por error.

57. "Si un grupo de la Tierra aterriza en un planeta, este evento es anunciado por los periódicos del mundo. No fue anunciado ningún otro aterrizaje en Nicoma, que no fuese el nuestro y el de los otros exploradores que nos precedieron."

¿Cuál de las siguientes afirmaciones es cierta, como consecuencia de lo anterior?

A. Si los periódicos anuncian un aterrizaje, es porque se ha producido.

B. Este grupo de seres es de la Tierra.

C. Este grupo de seres no es de la Tierra.

62. "¡Mira! Uno de nuestros exploradores saltó por una ventana y comenzó a huir. Dejó de correr, levantó sus brazos cuando un centinela gritó y le apuntó con la escopeta. Un grupo amistoso dejaría salir a sus huéspedes."

¿Cuál de las siguientes afirmaciones es cierta, como consecuencia de lo anterior?

A. Los grupos hostiles apresan a sus invitados.

B. Este grupo de seres es muy cuidadoso.

C. Este grupo de seres es hostil.

63. "Si hablásemos con nuestros exploradores descubriríamos, sin duda, si estos seres quieren negociar la paz. Podremos hablar con ellos si nos colamos, furtivamente, por la parte de atrás de la cárcel cuando los guardias cambian de posición."

¿Cuál de las siguientes afirmaciones es cierta, como consecuencia de lo anterior?

A. Podemos saber, con certeza, si estos seres quieren negociar la paz.

B. No podemos saber, con certeza, si estos seres harán la paz.

C. Nosotros no nos podemos colar en el cobertizo, si los centinelas son muy cuidadosos.

65. "Si les atacamos, mataremos algunos de ellos. Si matamos a algunos de ellos, perderemos información sobre Nicoma. Ahora no podemos permitirnos perder cualquier información sobre Nicoma."

¿Cuál de las siguientes afirmaciones es cierta, como consecuencia de lo anterior?

A. Debemos atacar.

B. Nosotros debemos matar a algunos de ellos.

C. No debemos atacar.

Parte II

Después de observar la aldea durante una hora, usted toma su grupo para volver al campamento. Ordena al sargento elaborar un informe para presentar al capitán. Para hacer el informe, el sargento toma como ciertas algunas ideas, aunque sin decirlo abiertamente. Estas ideas servirán de base para su razonamiento.

TED
Nº 53 • Primer semestre de 2023 • pp. 14-41
ISSN 2665-3184 impreso • ISSN 2323-0126 Web

Su trabajo ahora consiste en seleccionar las ideas que el sargento probablemente toma como supuestamente ciertas para justificar esos razonamientos, sin expresarlas realmente.

Aquí está un ejemplo:

66. "Los exploradores no pueden escapar debido a que no pueden derribar las paredes de la cabaña de piedra."

¿Cuál de las siguientes afirmaciones probablemente se da por supuesta?

A. Los Exploradores Pueden Saltar Por La Ventana.

B. Los Centinelas Están Alerta.

C. Todos los medios de escape son imposibles, excepto a través de las paredes.

Marque una respuesta.

La respuesta correcta es C. Entre todas las posibilidades, la C es la que más ayuda a la afirmación inicial. Marque C en su hoja de respuestas.

Para cada una de las siguientes preguntas hay una respuesta que puede ser considerada la mejor y más aceptable.

En esta parte de la historia también puede volver hacia una pregunta de atrás.

71. "Si al menos la mitad de los aldeanos son hombres, entonces tenemos que luchar contra la mitad, por lo menos."

¿Cuál de las siguientes afirmaciones probablemente se da por supuesta?

A. Las mujeres no son combatientes.

B. Los hombres son combatientes.

C. No podemos ganar si son todos los combatientes.

72. "No necesitaremos preocuparnos de más de diez a la vez, puesto que solamente hay diez pistolas."

¿Cuál de las siguientes afirmaciones probablemente se da por supuesta?

A. Las pistolas pueden hacernos daño.

B. Los cuchillos no hacen daño.

C. Solo las pistolas pueden hacernos daño.

73 "Solo tienen diez pistolas. Sé esto porque cada centinela tenía una y ocho más estaban apiladas en el centro del pueblo. Era todo lo que se podía ver."

¿Cuál de las siguientes afirmaciones probablemente se da por supuesta?

A. Todas las armas que tienen están a la vista.

B. No llevan armas bajo sus pieles de animales.

C. Las pistolas son su única arma de defensa.

74 "Los aldeanos no tienen centinelas en el exterior. Puedo garantizarlo porque miramos con mucha atención y no hemos visto ninguno."

¿Cuál de las siguientes afirmaciones probablemente se da por supuesta?

A. Los centinelas solo son empleados por personas que quieren que alguien vigile por ellas.

B. Los centinelas pueden ser vistos por personas atentas a ellas.

C. Si se viese un centinela, entonces esta no sería muy importante.

76. "Los aldeanos no son de la Tierra porque no hemos oído hablar de ningún otro aterrizaje en Nicoma proveniente de la Tierra."

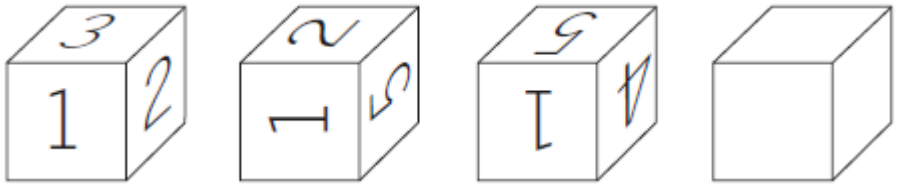
¿Cuál de las siguientes afirmaciones probablemente se da por supuesta?

A. Todos los aterrizajes en planetas son anunciados.

B. Todos los aterrizajes realizados por personas de la Tierra en otros planetas son anunciados a los otros exploradores terrestres.

C. Los exploradores de la Tierra no han oído hablar de aterrizajes hechos por exploradores de otros planetas.

26. Observa con atención los cambios de posición y orientación mostrados en las caras de los tres cubos primeros, como consecuencia del movimiento de rotación efectuado para pasar de uno a otro de izquierda a derecha.

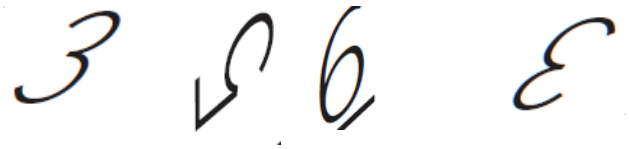


Escribe los números que deberían verse en las caras vacías frontal y lateral del cuarto cubo si se aplica una rotación más (atención a la orientación, arriba-abajo e izquierda-derecha que deben tener los tres números para continuar la secuencia correctamente).

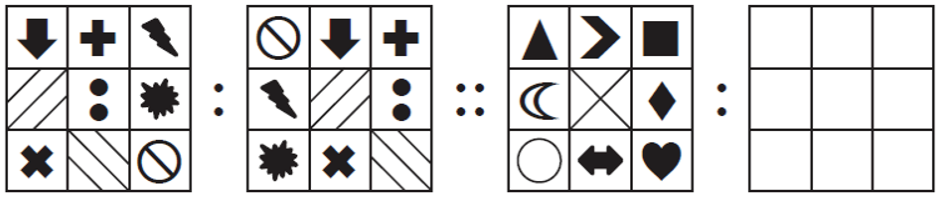
a. b. c. d.

¿Qué número debería verse en la cara vacía frontal del cuarto cubo si se aplica una rotación más?

¿Qué número debería verse en la cara vacía lateral derecha del cuarto cubo si se aplica una rotación más?



27. Observa cuidadosamente las secuencias de figuras contenidas en las nueve casillas de los dos primeros tableros, comparándolas, para intentar comprender la relación entre ellos.



A

B

Tu tarea consiste en señalar las imágenes que deberían contener las casillas A y B del cuarto tablero, si este se relacionase con el tercero de la misma manera que el segundo está relacionado con el primero.

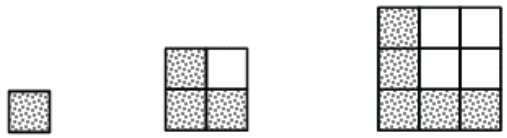
a. b. c. d.

¿Qué imagen debería tener la casilla A del cuarto tablero?

¿Qué imagen debería tener la casilla B del cuarto tablero?

Las figuras siguientes están construidas con unidades cuadradas y forman una secuencia regular de la cual se muestran las tres primeras etapas.

La etapa 3 de la secuencia está formada por nueve unidades cuadradas, de las cuales cinco están sombreadas y cuatro son blancas (sin sombra).



Etapa 1 Etapa 2 Etapa 3

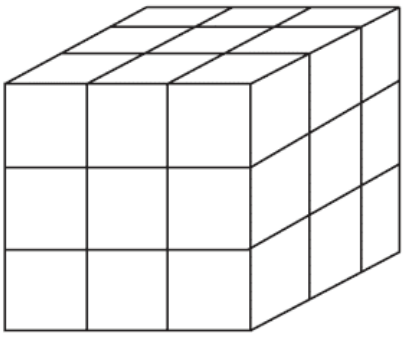
¿Cuántas unidades cuadradas blancas tendrá la figura de la etapa número once?

_____

¿Cuántas unidades cuadradas sombreadas tendrá la figura de la etapa número catorce?

_____

El cubo de la figura está formado por 3 x 3 x 3 cubitos más pequeños e iguales; el cubo se cayó en una gran lata de pintura roja, quedando todas sus caras sin excepción pintadas de rojo.



Responde las siguientes preguntas:

1. ¿Cuántos cubitos pequeños tienen solo dos caras pintadas de rojo?
2. ¿Cuántos cubitos pequeños tienen solo tres caras pintadas de rojo?
3. ¿Cuántos cubitos pequeños no tienen ninguna cara pintada de rojo?

En la cuestión siguiente tu tarea consiste primeramente en leer con detenimiento y atención la información proporcionada en el texto inicial. Después, para cada una de las afirmaciones resaltadas en negrita y teniendo en cuenta la información proporcionada en el texto inicial, **señala el grado de veracidad que crees que tiene cada decisión** (marca uno de los cinco grados de respuesta que se ofrecen, desde Verdadera hasta Falsa).

Texto de información

Investigadores de una Universidad muy respetada informaron en un artículo de una revista científica que el maíz genéticamente modificado (GM) vendido por GenetiOrg puede representar una amenaza para la salud humana. Los investigadores aislaron genes provenientes del maíz genéticamente modificado, dentro de las células de los pollos criados en las granjas cercanas. El estudio fue financiado principalmente por la Agencia Nacional para la Ciencia del Gobierno, aunque una pequeña porción de la financiación fue proporcionada por Defensa de la Naturaleza. Defensa de la

Naturaleza es una organización ambiental privada que ese opone a la utilización de cultivos transgénicos. Poco después de que estos resultados de la investigación fuesen conocidos por el público en general, el precio de las acciones de GenetiOrg cayó 13 puntos.

a. Verdadera
b. Probablemente cierta
c. Datos insuficientes
d. Probablemente falsa
e. Falsa

Decisión	a.	b.	c.	d.	e.
(3) 12122 33. El estudio demostró que los cultivos genéticamente modificados son nocivos como alimento de los seres humanos.					
(4) 12123 34. Los investigadores de la Universidad inventaron sus resultados para mantener sus fondos de investigación del grupo ambiental.					
(5) 12124 35. Defensa de la Naturaleza no utilizará los resultados del estudio para dar apoyo a su posición contra los cultivos genéticamente modificados.					