¿Puede ser evaluado el pensamiento crítico de forma breve?

Carlos Saiz[1]
Leandro S. Almeida[2]
Silvia F. Rivas[1]

[1]*Universidad de Salamanca, Salamanca, Espanha*
[2]*Universidade do Minho, Braga, Portugal*

Resumen
El objetivo de la investigación ha sido validar una versión breve de la prueba completa de pensamiento crítico (PC) PENCRISAL, en español y portugués. Aquí se presenta la primera. Esta versión breve consta de 2 factores, con 3 ítems cada uno. Un factor evalúa argumentación general, y el otro las formas de razonamiento más empleadas a diario. Además, ambos factores evalúan indirectamente la toma de decisiones y la solución de problemas, gracias a la naturaleza de los ítems, en los que se plantean problemas cotidianos que se deben resolver y, en ocasiones, hay que tomar decisiones. El análisis factorial confirmatorio nos ofrece índices de ajuste adecuados que avalan la estructura de la versión breve del test presentado. Los coeficientes de fiabilidad y validez son razonablemente robustos, lo que hace que esta prueba sirva a los objetivos de investigación establecidos.
Palabras clave: pensamiento crítico; evaluación; enseñanza superior.

Pode o pensamento crítico ser avaliado de forma breve?

Resumo
O objetivo do nosso estudo foi validar uma versão breve do Teste Completo de Pensamento Crítico (PC) PENCRISAL, em espanhol e português. Aqui apresentamos o primeiro. Esta versão resumida consiste em 2 fatores, com 3 itens cada. Um fator avalia a argumentação geral e o outro as formas de raciocínio mais utilizadas no dia a dia. Além disso, ambos os fatores avaliam indiretamente a tomada de decisão e a resolução de problemas, graças à natureza dos itens, nos quais se levantam problemas diários que devem ser resolvidos e para os quais, às vezes, devem ser tomadas decisões. A análise fatorial confirmatória nos oferece índices adequados de ajustamento que sustentam a estrutura da versão resumida do teste apresentado. Os coeficientes de confiabilidade e validade são razoavelmente robustos, o que faz com que este teste sirva para os objetivos de investigação declarados.
Palavras-chave: pensamento crítico; avaliação; ensino superior.

Can critical thinking be briefly assessed?

Abstract
The aim of our study was to validate a short version of the PENCRISAL complete critical thinking test, in Spanish and Portuguese. Here we present the first. This short version consists of 2 factors, with 3 items each. One factor assesses general argumentation, and the other the forms of reasoning most used on a daily basis. In addition, both factors indirectly evaluate decision-making and problem solving, thanks to the nature of the items, in which daily problems are raised that must be solved and for which decisions must sometimes be made. The confirmatory factor analysis offers us solid and consistent indices that support the structure of the short version of the test presented. The reliability and validity coefficients are reasonably robust, as to make this test serve very well for the stated objectives.
Keywords: critical thinking; evaluation; higher education.

Entre las habilidades cognitivas relevantes en la sociedad del conocimiento de hoy día destaca el pensamiento crítico (PC). Estas competencias permiten a las personas ser autónomas, lo que se traduce por ser capaz de tomar decisiones y resolver problemas de manera eficaz. En este sentido, al hablar de la calidad de la educación, se proponen cada vez, con más frecuencia, sistemas educativos que sean capaces de desarrollar tales competencias en los estudiantes (Heard et al., 2020; Rivas et al., 2020; Saiz et al., 2020; Uribe et al., 2017).

El PC se encuentra ya en los orígenes de la filosofía; sin embargo, podemos situar su nacimiento moderno en el trabajo de Robert Ennis de 1956 (Ennis, 1956), en el que desarrolla su primera concepción sobre el PC, para consolidarla más tarde en su libro ampliamente referenciado cuatro décadas después (Ennis, 1996). El trabajo de Ennis de 1956 nos interesa más aquí porque es la investigación que va a dar lugar a la primera prueba estandarizada de PC. El desarrollo de este instrumento es el objeto de su tesis doctoral que se publica tres años más tarde basándose en esas primeras reflexiones sobre lo que él entiende por PC (Ennis, 1959). Después vendrán otros testes, por parte de este autor, como el popular test *Cornell* (Ennis et al., 1985).

http://dx.doi.org/10.1590/1413-8271202126nesp13

Este es el punto de partida de la respuesta que daremos a la pregunta planteada en este trabajo, y la historia sobre la evaluación del PC que mencionaremos solo será la necesaria para justificar lo mejor posible nuestra propuesta de evaluación.

Esa primera prueba de Ennis fue una apuesta novedosa y arriesgada. En aquellos años, el intentar medir procesos cognitivos de tal complejidad era un proyecto para el que aún ni había suficiente conocimiento ni sensibilidad científica; esto segundo, entre otras razones, por una fundamental, porque competía directamente con la medida de la inteligencia que se encontraba en su máximo esplendor, de la mano del cociente intelectual (CI). Sin embargo, Ennis logró desarrollar esta prueba y otras, inaugurando una trayectoria muy fructífera, gracias al aprovechamiento que hizo de la tradición psicométrica muy influyente por aquel entonces. Este contexto reforzaba las medidas cuantitativas exclusivamente, algo que llevó quizás al desarrollo de pruebas de formato cerrado, principalmente de elección múltiple. La influencia positiva de este entorno psicométrico sobre la objetividad y el rigor de medida tiene un precio en el ámbito de los procesos cognitivos superiores, a saber, su mirada limitada.

Las pruebas objetivas y cerradas ofrecen una gran ventaja respecto a la consecución de la fiabilidad de la medida, pues la garantizan con bastante facilidad. Medir bien lo que se quiere medir se asegura mejor con un formato cerrado que con otros. No obstante, medir lo que se dice que uno mide, no tanto. La validez siempre es muy escurridiza cuando se trata de la cognición de orden superior, como es todo lo que podamos entender por argumentación, explicación, pensamiento profundo o crítico (Wechsler et al., 2018). Un formato de respuesta cerrado nos devuelve unas marcas, una elección de una opción de entre otras simples y explícitas. La comparación que mejor nos permite entender esa limitación es recurrir a las dos formas clásicas de medir la memoria: tareas de reconocimiento y tareas de recuerdo. En las primeras, simplemente realizamos una labor de discriminación entre información que se nos da, mientras que, en las segundas, debemos proporcionar esa información. Claramente, entendemos que la riqueza de las respuestas de las tareas de recuerdo es muy superior a las de reconocimiento; en ellas, podemos descubrir buena parte de los mecanismos de codificación y recuperación de nuestra memoria. Lo mismo nos sucede con las respuestas de elección múltiple en las pruebas de PC, nos ofrecen muy poco o nada de los pasos o andadura mental que recorre quien las contesta; esto es una limitación grave para saber realmente qué es lo que estamos midiendo.

Decir que mido PC solo por las marcas de opción que nos devuelve quien responde a la prueba es demasiado ingenuo. La validez nos exige demostrar qué proceso mental está aconteciendo cuando se elige una opción de las que se ofrecen, y esto, no se nos ocurre cómo se puede saber con solo marcar esa opción. Las respuestas que se dan pueden ser fruto de diferentes estrategias o mecanismos, pero no tenemos forma de averiguar cuál es de esos posible, y lo más grave, de saber si son de pensamiento y no de otra cosa. Sin esta certeza ¿cómo podemos hablar de que medimos lo que decimos? ¿Cómo asegurar que la prueba es válida? Con esta clase de pruebas, no es posible. Sí, cierto que las estructuras factoriales muestran índices adecuados con las dimensiones propuestas, pero esas dimensiones ¿son lo que dicen que son? No es nuestro objetivo aquí echar por la borda todo el trabajo metodológico realizado y bien fundamentado en este ámbito, obviamente. Solo queremos llamar la atención sobre el hecho de que esta forma de evaluar no nos da todo lo que necesitamos para entender y medir bien procesos como los de argumentar, explicar, decidir o resolver. Tampoco queremos negar otras formas de diseñar pruebas cerradas que sí nos permiten saber con certeza qué proceso acontece frente a cada ítem o cuestión que se plantee. Pero esta es otra historia que estamos desarrollando y que verá la luz a su debido tiempo. Lo que ha sucedido hasta ahora es que las pruebas de formato cerrado que se han creado son demasiado limitadas como para poder captar inequívocamente lo que dicen que miden. Aunque circunscrito solo a los mecanismos de argumentación, Govier (1987) señalaba las mismas limitaciones y gravedad del problema de la validez.

Si estos instrumentos no nos permiten captar realmente los procesos que empleamos al afrontarlos ¿qué tipo de pruebas podemos emplear en su lugar? Han tenido que pasar algunas décadas para que una experta y referente en este campo diera una respuesta convincente a esta cuestión. Halpern (2003) fue consciente de este problema y comenzó a desarrollar por aquel entonces una prueba que lo solucionara, y de la que se publicó una primera versión en 2006 (Nieto & Saiz, 2008), y la versión definitiva un poco después (Halpern, 2012). Esta prueba posee varias virtudes que conviene mencionar. El test *Halpern Critical Thinking Assessment* (HCTA) es de formato mixto: abierto y cerrado, sus ítems son problemáticas cotidianas, y su sistema de puntuación en

la versión cerrada es sencillo y claro (Butler et al., 2017; Halpern & Dunn, 2021). El formato abierto permite solucionar el problema de validez mencionado antes al permitir a quien realiza el test explicar por qué responde del modo que lo hace. De este modo, podemos averiguar qué mecanismos o procesos está empleando para dar su respuesta, de forma que nos aseguramos de si está procediendo como se espera que lo haga o, por el contrario, elige un modo diferente para responder. De esta manera, podemos saber si contesta, por ejemplo, a los ítems de argumentación con los procesos correspondientes, y no con otros, con independencia del acierto o error de su respuesta. Esto sí nos permite asegurar la validez con rigor.

La segunda peculiaridad de la prueba, y que es una ingeniosa e importante aportación, consiste en emplear problemáticas cotidianas, como ítems del test. Este rasgo contribuye a que la prueba sea enormemente ecológica, o cercana para quien responde, pues se le pide que afronte esta clase de situaciones para las que se encuentra muy familiarizado y le resultan muy cercanas. Esto hace que la prueba sea muy interesante y amigable, algo muy deseable siempre por altamente motivante. El emplear situaciones-problema tiene una relevancia fundamental, que debemos enfatizar, para que no pase desapercibida su trascendencia, en comparación a lo que se ha venido haciendo hasta ahora. Démonos cuenta de que hay demasiadas pruebas de PC en las que los ítems son autoinformes, o petición de valoraciones o calibraciones que se solicitan como respuesta a los participantes, algo de escaso valor a la hora de evaluar esta clase de competencias. Pero, además, cuando esto no es así, lo que los ítems de esas pruebas plantean son problemas escasamente relevantes o acertijos triviales, para los que difícilmente necesitaremos responder empleando los procesos cognitivos que pretendemos medir.

Sin embargo, cuando utilizamos problemáticas cotidianas presentamos situaciones relevantes que hay que resolver empleando lo que corresponda en cada ocasión, es decir, se nos pide que digamos cómo actuaríamos en esas circunstancias, qué haríamos ahí. Percatémonos de que estamos pidiendo conductas, acciones, no pareceres, ni impresiones, ni valoraciones. Es necesario alejarnos de creencias expresadas y acercarnos solo a comportamientos, a hechos contrastables. Si en una situación cotidiana se propone una solución que es la única posible como la mejor, nos aseguramos de que quien la afronte está entendiendo y resolviendo según lo esperado, y no valorando diferentes opciones que puedan ser aceptables. Lo que esta clase de ítems capta, y en esto descansa su trascendencia, son resultados, los cambios que habría que realizar para solucionar la situación (Saiz, 2020). En realidad, la naturaleza de este instrumento aporta algo insustituible por imprescindible: enfrenta al que realiza la prueba a realidades relevantes para las que hay que ofrecer conductas que la modifiquen o la cambien. No hay posibilidad de especular, de ofrecer pareceres, sino solo de resolver con acciones. Para nosotros, es esencial esta forma de evaluar por resultados, por conductas que cambien la realidad y resuelvan realmente los problemas que importan. Y esto sí se logra con esta clase de problemáticas y formato de respuestas abiertas.

En su momento, contemplamos el HCTA como una prueba idónea para la evaluación del PC, y desarrollamos un proyecto encaminado a su validación en español. Sin embargo, fracasamos en nuestro objetivo (Nieto et al., 2009). Al no poder disponer de un test adecuado en nuestro idioma para medir el PC, estudiamos con mayor profundidad la prueba HCTA. Seguimos convencidos de sus excelencias, como el formato abierto y, sobre todo, la naturaleza de los ítems, tal como acabamos de justificar. Sin embargo, encontramos una debilidad importante: la forma de corrección de algunas dimensiones. Por ejemplo, algunos ítems de solución de problemas se puntuaban como correctos, con solo indicar que se iba a buscar información para solucionar la situación, pero no por emplear una estrategia de solución adecuada. Al proceder así en la corrección, se pone en entredicho la validez de algunas dimensiones, pues se da como bueno aquello que no es lo pertinente para lo que se demanda en esa problemática determinada. En su día, publicamos esta revisión de la prueba y nuestra propuesta de solución a esta limitación (Saiz & Rivas, 2008). Nuestra forma de resolver esta importante dificultad consistió en emplear una vieja metodología muy olvidada, el *análisis de tareas* (Donders, 1868/1969). Brevemente, lo que Donders propone es realmente ingenioso: utilizar tareas concretas para poner en marcha procesos concretos. En toda tarea de Donders, se fijan los estímulos que se deben presentar y las respuestas que se deben emitir. En HCTA, se determina solo lo primero, esto es, las situaciones-problema, pero no lo segundo, las respuestas correctas para cada situación; por esto, no podemos saber cuál es la estrategia adecuada para responder bien, y tampoco cómo corregir las respuestas de modo consistente (los pormenores de estos análisis se encuentran en Saiz & Rivas, 2008).

A raíz de estos dos problemas, el fracaso en la validación del test de Halpern y de los problemas de validez indicados, tomamos la decisión de construir una prueba para evaluar el PC. Este instrumento conservó la gran aportación del HCTA, a saber, emplear ítems que sean problemáticas cotidianas y, además, optar solo por un formato de respuesta abierto. Se construyeron y seleccionaron situaciones problemas que fueran adecuados a nuestras características culturales, con el fin de lograr familiaridad y motivación. Este proyecto comenzó en Saiz y Rivas (2008), y se publicó la validación de la prueba, llamada PENCRISAL, en Rivas y Saiz (2012). Adicionalmente, se realizó una adaptación peruana en Rivas et al. (2014). El PENCRISAL consta de 35 ítems o problemáticas cotidianas, de formato abierto, que se corrigen de un modo sencillo, puntuando 0, 1 y 2: incorrecta, correcta, pero sin adecuada justificación, y correcta. Estos ítems se estructuran en 5 dimensiones: *deducción, inducción, argumentación o razonamiento práctico, toma de decisiones, y solución de problemas*. Cada una de las dimensiones está compuesta por 7 ítems (para más detalles, ver Rivas & Saiz, 2012). Del mismo modo que sucede con la mayoría de las pruebas de evaluación del PC, en esta se encuentran las dimensiones más consensuadas y aceptadas del PC. De hecho, Ennis (2003) revisa las 21 pruebas existentes por aquel entonces y en todas ellas aparecen estas dimensiones que son las realmente representativas de este complejo constructo cognitivo.

El test PENCRISAL, actualmente, es la única prueba de estas características en español. Desde su validación, se ha aplicado a cientos de participantes en diferentes países de Iberoamérica. Este instrumento es adecuado para un diagnóstico completo del PC, al poder evaluar cinco factores con un número de ítems que recogen la mayoría de los procesos implicados en cada uno ellos. Y así ha funcionado y sigue funcionando. Por lo tanto, parece cubierta esta necesidad de evaluación de un modo conveniente. Podemos decir que sí, que se cubren todos los objetivos de medida que se puedan plantear. Sin embargo, siempre lo extenso tiene su inconveniente en la duración. Esta prueba necesita de entre 90 y 115 minutos de media para completarse. En estos tiempos tan "líquidos", conseguir datos o muestra, cada vez se hace más difícil. Por esta razón, en este trabajo se plantea el objetivo de conseguir disponer de una prueba reducida para la evaluación del PC. La finalidad de este proyecto es fundamentalmente práctica, disponer de una medida más corta del PC, que nos permita evaluar con más rapidez a muestras más amplias en proyectos que solo necesiten de un indicador global del nivel de PC, y que no busquen efectuar diagnósticos pormenorizados del mismo. De este modo, podemos obtener con relativa facilidad medidas de poblaciones difíciles de conseguir con pruebas cuya duración sea muy extensa. Para nuestros fines, llevamos un tiempo estudiando y aplicando una versión reducida del PENCRISAL en dos idiomas, en español y en portugués. En este trabajo solo podemos adelantar la versión en nuestro idioma, dado que hemos encontrado diferencias en la validación en ambos idiomas. Por lo tanto, la prueba que hemos conseguido validar es la versión abreviada en español. La versión portuguesa será objeto de un trabajo futuro.

La primera versión abreviada con la que hemos trabajado constaba de 20 ítems de esos 35, 4 ítems por dimensión. Los criterios de selección empleados fueron: a) representatividad de los ítems en cada dimensión, b) familiaridad, y c) nivel de dificultad. Con la experiencia acumulada de multitud de aplicaciones de la versión completa del test, podemos saber qué ítems mantienen un mejor equilibrio por dimensión, cuáles son más cercanos a las situaciones cotidianas de las personas, y cuáles más asequibles (Rivas & Saiz, 2012; Rivas et al., 2014; Saiz & Rivas, 2008). Como es lógico, es importante que los ítems seleccionados sean la mejor muestra de cada dimensión, esos cuatro elegidos deben ser los más relevantes dentro de cada una de ellas. Además, los ítems seleccionados deben recoger problemáticas diarias familiares y frecuentes, con el fin de lograr una mejor activación de los procesos cognitivos implicados. Por último, fruto de nuestra experiencia, hemos visto que el test completo es de una dificultad muy alta, lo que hace que la variabilidad de las respuestas sea baja, de modo que puede dificultar captar ciertos mecanismos de pensamiento por un efecto de fatiga de los sujetos. Por ello, se ha optado por seleccionar también los ítems de menor dificultad de respuesta, dado que la prueba posee un cierto efecto de suelo. No fue fácil esta selección debido a la interacción de los tres criterios. Por lo tanto, hemos seleccionado los ítems lo mejor que pudimos de acuerdo con esos criterios.

Sin embargo, a pesar del esfuerzo y de aplicar esta versión reducida a una muestra de más de 300 participantes, no hemos logrado validarla. La estructura factorial no correspondía con la de la prueba completa, con sus 35 ítems. Después de un análisis detenido de los protocolos de respuesta, hemos visto que las puntuaciones de los ítems de las dimensiones de toma de decisiones (TD) y solución de problemas (SP) eran

muy inestables, dificultando la consistencia deseada en la relación que debería existir entre los ítems de cada dimensión. Sabemos por la experiencia de años de aplicación que las problemáticas de TD y SP generales son difíciles de acertar, por ser los ítems más abiertos y menos vinculados a estrategias concretas. En los ítems específicos de estas dimensiones, como los referentes a heurísticos, estrategias medio-fin, o de regularidades, no había este problema; por el contrario, los ítems de las diferentes dimensiones de razonamiento se manifestaban razonablemente consistente, fueran cuales fueran los tipos de argumentación. Esto se puede comprender bien, porque las problemáticas de argumentación en general siempre están mejor estructuradas, de modo que las respuestas resultan menos dispersas. Las problemáticas de TD y SP son más borrosas y difusas que las de argumentación. Por lo tanto, la consistencia de las respuestas también. Como sabemos, esta dispersión inestable de los resultados dificulta la obtención de unos criterios de validación psicométrica razonables, y las estructuras factoriales resultantes no sustentan la validez de constructo (Kretzschmar et al., 2016).

De nuevo, después de un estudio detenido de estos datos, hemos visto que una versión abreviada de 20 ítems es difícil de obtener; por lo tanto, hemos pasado a indagar la posibilidad de una versión más corta de la prueba, esto es, hemos trabajado con una *versión más breve*, en lugar de con una abreviada. En esta parte del estudio, hemos buscados una configuración que fuera conceptualmente coherente y psicométricamente consistente. No olvidemos, que nuestro propósito es obtener una versión reducida del test, con el fin de poder evaluar salvando las dificultades de la longitud de la prueba, logrando que sea asequible por tiempo, familiaridad y dificultad, y también por su fácil administración en muestras amplias. Todo esto, creemos, permitirá una medida representativa y manejable del PC.

Método

Participantes

Para la validación de esta versión abreviada, hemos empleado una muestra de 340 participantes, estudiantes de primero de la Universidad de Salamanca. La muestra estaba formada por los cursos completos que se utilizaron para responder a la prueba. De estos 340 participantes, la mayoría son mujeres (82.7%). Las edades oscilan entre 17 y 35, siendo la media de 19.26 años (desviación típica de 2.78).

Instrumento

Como hemos señalado, buscamos una versión breve del PENCRISAL. No hemos logrado una versión abreviada del mismo por las razones ya aducidas, de modo que hemos explorado varias posibilidades consistentes en la combinación de varios ítems dentro de cada factor. Hemos encontrado agrupamientos de algunos ítems en dos factores con algunas incongruencias conceptuales, en comparación con las cinco dimensiones del test de la versión abreviada (20 ítems). Sin embargo, finalmente, la estructura factorial consistente que hemos obtenido ha sido una de 6 ítems agrupados en 2 factores o dimensiones, en concreto: (i) El factor 1 (razonamiento general -RGRAL-) está formado por el ítem 4 de razonamiento proposicional (RPRO), el ítem 14 de razonamiento analógico (RANLG), y el ítem 18 de argumentación o razonamiento práctico (RPRA); (ii) el factor 2 (razonamiento práctico -RPRA-) está formado por el ítem 3, argumentación (RPRA), el ítem 19, falacia (RPRA), y el ítem 20, falacia (RPRA). Ver tabla 1.

El factor 2 (f2) es el más homogéneo (ver tabla 1) pues los tres ítems son de argumentación, uno de argumentación (i3) y dos de falacias (i19, i20) o argumentación infundada. La argumentación es la competencia de razonamiento práctico más fácil de entender y una

Tabla 1.
Distribución de Ítems y Factores

Ítems	Factor 1: RGRAL	Factor 2: RPRA
3		Argumentación (RPRA)
4	Razonamiento proposicional (RPRO)	
14	Razonamiento analógico (RANLG)	
18	Argumentación (RPRA)	
19		Falacia (RPRA)
20		Falacia (RPRA)

de las pocas que se adquiere antes de los 16 años. Las falacias son sutiles falsos razonamientos, difíciles de manejar, y no suelen dominarse antes de esa edad de 16 años en la que se consolidan las operaciones formales piagetianas. Así, este factor (f2) evalúa el razonamiento práctico de forma completa, ya que afronta la correcta e incorrecta argumentación (su estructura y valoración). Por lo tanto, dicho factor cubriría las competencias generales de razonamiento no vinculadas a formas específicas de inferencia, como veremos en el factor 1. Debemos recordar que todos los ítems son problemas cotidianos, frente a los cuales debemos decidir qué hacer para resolverlos. Aunque no se miden estrategias específicas de decisión y solución de problemas, sí se deben resolver como tales, recurriendo a los mecanismos de inferencia planteados en cada ítem.

El factor 1 (f1) es interesante por otras razones (ver tabla 1); consta de un ítem de razonamiento proposicional (i4), otro de razonamiento analógico (i14), y el último de argumentación (i18). Este factor contempla una forma de razonamiento formal, otra metafórica, y una general, de nuevo. Son tres formas de inferencia representativas de nuestro funcionamiento diario. La formal es la más sencilla y común de esta clase. Las analogías, el "es como si…", son inferencias muy frecuentes, para cuando no podemos ser especialmente precisos; finalmente, la argumentación, ya lo hemos dicho, incluye modos globales de inferencias difíciles de formalizar en algunas situaciones. Los tres ítems contemplan y evalúan formas de inferencias frecuentes y representativas de nuestro funcionamiento diario. Recordemos que, igual que en el factor 2, también aquí las problemáticas planteadas en los ítems deben resolverse decidiendo o proponiendo cursos de acción.

Este conjunto de 6 ítems permite medir competencias esenciales del PC, como son las formas de inferencia más representativas y, a la vez, el formato de situación-problema de los ítems permite recoger la forma de decidir y resolver en general (un ejemplo de los ítems del test, se puede encontrar en Rivas & Saiz, 2012, en la página 22). Bien es verdad que no podemos evaluar estrategias específicas de TD y SP. Una versión breve necesariamente sacrifica competencias que no pueden contemplarse con tan pocos ítems; sin embargo, conserva las competencias generales y más representativas del PC.

Procedimientos

Los estudiantes han realizado la prueba de forma libre y voluntaria, esto es, con consentimiento informado, y atendiendo a las recomendaciones deontológicas que la universidad pide. El tiempo de aplicación fue de una semana, durante el cual podían cumplimentar el test en el momento que consideraran oportuno y en el lugar que mejor les conviniera. Al ser una prueba de potencia, las instrucciones eran que podían emplear el tiempo que necesitaran, pues lo importante era la calidad de las respuestas, no su rapidez. La participación en la prueba permitía a los estudiantes subir unas décimas en su nota final (0.25 puntos, sobre 10), lo que conocían con mucho tiempo de antelación y antes del comienzo de su participación. La aplicación se realizó a través de internet, desde la plataforma Selectsurvey. NETv5.0, en https://24.selectsurvey.net/pensamiento-critico/Login.aspx

Para confirmar la dimensionalidad de los ítems seleccionados para esta versión breve de la prueba, se recurrió al análisis factorial confirmatorio mediante el programa M-Plus (Versión 8.6; Muthén & Muthén, 2019). Dada la métrica básicamente dicotómica de los ítems (valores entre 0 y 2), utilizamos el estimador de *media y varianza de mínimos cuadrados ponderados* (WLSMV). Como criterios para ajustar el modelo, utilizamos el *Chi-cuadrado* ($\chi 2$), el *índice de ajuste comparativo* (CFI), *el índice de Tucker-Lewis* (TLI), la *raíz del error cuadrático medio de aproximación* (RMSEA) *y la raíz cuadrática media residual estandarizada* (SRMR). Según la literatura, buscamos un valor de Chi-cuadrado (χ^2) menor a 3.0, valores CFI y TLI mayores a .90 (Hu & Bentler, 1999). Finalmente, un valor inferior a .08, para SRMR, o entre .06 y .08 para RMSEA, sugieren un buen índice de ajuste (MacCallum et al., 1996).

Resultados

En primer lugar, hemos analizado la estructura factorial de los seis ítems distribuidos en los dos factores: 1) razonamiento general, y 2) razonamiento práctico. Los datos del análisis factorial confirmatorio producían adecuados índices de ajuste: $\chi^2 = 4.31$; CFI = .99; TLI = .98; RMSEA = .02; y WRMR = .51. En figura 1 presentamos los *loadings* obtenidos.

Como podemos ver en la figura 1, los ítems saturan satisfactoriamente en ambos factores (índices por encima de .40). Además, se observa una correlación elevada entre los dos factores identificados de razonamiento general y razonamiento práctico (un valor superior a .60)

Tomando las puntuaciones de los estudiantes, en la tabla 2, presentamos los índices descriptivos, teniendo

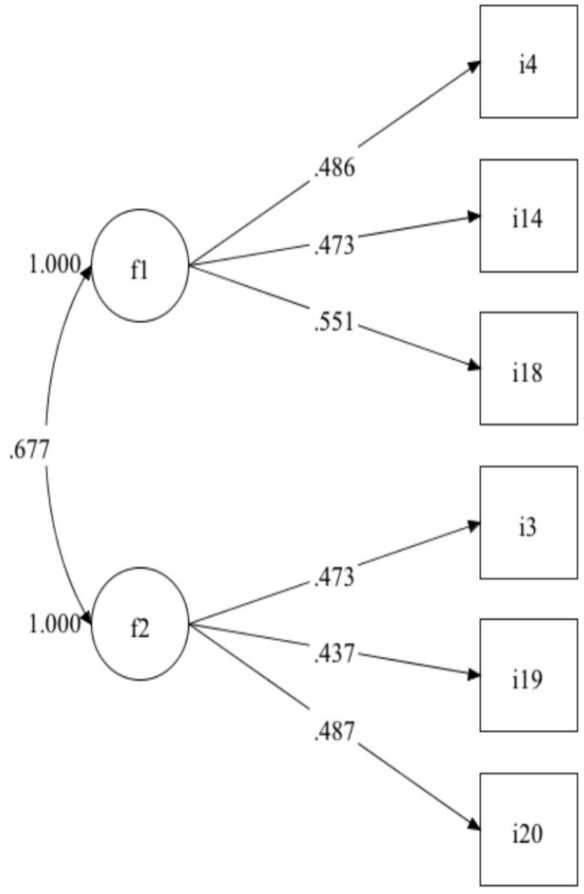

Figura 1. Estructura Factorial

en cuenta la muestra total, y ésta también en función del sexo y de la edad; esta última variable la hemos dividido en dos: estudiantes de menos edad (hasta 19 años) y de más edad (desde 20 años). Se ofrecen las medias y las desviaciones típicas, junto con los valores mínimos y máximos, así como los coeficientes de asimetría y curtosis de la distribución.

Los valores obtenidos muestran que siempre se da el caso de que algún alumno obtiene la puntuación mínima posible (cero puntos) y también la máxima posible (seis puntos). Al mismo tiempo, los índices de asimetría y de curtosis son bajos, en correspondencias con los presupuestos de normalidad gaussiana de las puntuaciones. Las puntuaciones de los estudiantes en la dimensión razonamiento práctico (ítems 3, 19, 20) tienden a ser siempre superiores a la obtenida en la dimensión razonamiento general (ítems 4, 14, 18). Analizando las medias en función del sexo y de la edad, se observa que las mejores puntuaciones se encuentran en los estudiantes masculinos y en el grupo de menor edad (hasta 19 años). Estas diferencias en función del sexo son estadísticamente significativas para la primera dimensión (t [1, 327] = 2.452, p = .015), siendo casi estadísticamente significativas para el conjunto de las dos dimensiones o seis ítems (t [1, 327] = 1.835, p = .067); sin embargo, no son estadísticamente significativas en la segunda dimensión (t [1, 327] = .640, p = .522).

Tabla 2.
Estadística Descriptiva de los Resultados

Muestra	Factores	Min.-Max.	M	DT	Asimetría	Curtosis
Global	1	0 - 5	2.25	1.22	.024	-.898
	2	0 - 6	2.40	1.48	.254	-.674
	1+2	0 - 11	4.65	2.14	.168	-.417
Masculino	1	0 - 5	2.33	1.21	-.054	-.812
	2	0 - 6	2.42	1.51	.256	-.707
	1+2	0 - 11	4.75	2.13	.162	-.405
Femenino	1	0 - 4	1.89	1.23	.443	-.882
	2	0 - 5	2.28	1.33	.164	-.618
	1+2	0 - 9	4.18	2.16	.230	-.401
Menos edad	1	0 - 5	2.36	1.21	-.028	-.918
	2	0 - 6	2.47	1.41	.193	-.611
	1+2	0 - 11	4.83	2.03	.169	-.389
Más edad	1	0 - 5	1.91	1.21	.218	-.723
	2	0 - 6	2.17	1.69	.503	-.746
	1+2	0 - 11	4.08	2.40	.396	-.411

A nivel de edad, los estudiantes más jóvenes muestran mejores resultados, con una diferencia estadísticamente significativa en las medias en la primera dimensión (t [1, 327] = 2.88, p = .004), y en ambas conjuntamente (t [1, 327] = 2.745, p = .006); pero no así en la segunda dimensión, la cual no sale significativa estadísticamente (t [1, 327] = 1.590, p = .113).

Conclusiones

El *test breve* de PC que hemos presentado y validado es novedoso porque permite una evaluación del nivel general del PC de forma rápida y a poblaciones amplias, algo que es necesario en la mayoría de las investigaciones del campo, y para lo cual no se disponía hasta ahora de una prueba de estas características en español. El poder establecer un nivel de PC antes y después de cualquier intervención destinada a mejorar esas competencias, por ejemplo, nos permite realizar mucha más investigación aplicada de la que se viene realizando, al reducir los problemas de obtención de muestra y el tiempo de evaluación de la misma.

Los factores de la prueba que se han obtenido cubren razonablemente bien los procesos cognitivos más relevantes, pues abarcan todo lo referente a las diferentes formas de razonamiento y argumentación, y contempla indirectamente las destrezas que tienen que ver con toma de decisiones y solución de problemas. Estas dimensiones hacen que la prueba sea susceptible de emplearse en la mayoría de los estudios sobre esta clase de competencias transversales, con una excepción. Este test no es adecuado emplearlo con fines de diagnóstico pormenorizado, donde lo que se necesita es disponer de indicadores del nivel en cada una de las habilidades específicas de PC que posee cada participante. Al buscar un barrido rápido del nivel de competencia general de PC, se han sacrificado muchos ítems encargados de medir, por ejemplo, estrategias específicas de TD o de SP. Con esta prueba esto no es posible. Para ello, debemos emplear la versión completa del test.

Como se ha podido observar en el apartado de resultados, esta prueba posee una validez de constructo sólida, tal cual reflejan los pesos factoriales de la figura 1 (Kretzschmar et al., 2016). Su estructura factorial es consistente teóricamente y con sentido. El agrupamiento de los ítems en los dos factores descritos organiza con coherencia la mayoría de las inferencias relevantes y frecuentes que utilizamos. Esto permite una valoración precisa y rápida de los diferentes tipos de juicios que empleamos en nuestra vida diaria, a la hora de afrontar situaciones y problemas importantes. El poder establecer niveles o categorías, por otro lado, nos brinda la oportunidad de indagar nuevas relaciones entre el PC y otros componentes no cognitivos relevantes, como lo motivacional o disposicional.

El siguiente avance de este proyecto, como hemos dicho, es validar esta versión breve en portugués. Hemos encontrado disparidad entre ambas versiones (española y portuguesa) que dificultan este objetivo, quizás por las diferencias culturales de los ítems. Esto nos obliga a efectuar una selección distinta de las problemáticas teniendo en cuenta estas singularidades.

Finalmente, estamos buscando el modo de desarrollar un instrumento de evaluación que recoja las virtudes de la versión breve y completa a la vez, como también ya hemos apuntado antes. Sin embargo, recoger lo bueno de estos dos mundos nos exige un proyecto más a largo plazo que, esperamos, poder terminar con éxito.

Referencias

Butler, H. A., Pentoney, C., & Bong, M. P. (2017). Predicting real-world outcomes: Critical thinking ability is a better predictor of life decisions than intelligence. *Thinking Skills and Creativity, 25*, 38-46. https://doi.org/10.1016/j.tsc.2017.06.005.

Donders, F. C. (1969). On the speed of mental processes. *Acta Psychologica*, *30*, 412-431. Original: 1868.

Ennis, R. H. (1956). Critical thinking: More on its motivation. *Progressive Education*, *33*, 75-78.

Ennis, R. H. (1959). *The development of a critical thinking test.* Unpublished doctoral dissertation, University of Illinois. University Microfilms #59-00505.

Ennis, R. H. (1996). *Critical thinking*. Prentice-Hall.

Ennis, R. H. (2003). Critical thinking assessment. En D. Fasko (Ed.*), Critical thinking and reasoning. Current research, theory, and practice* (pp. 293–313). Hampton Press.

Ennis, R. H., Millman, J., & Tomko, T.N. (1985). *Cornell Critical Thinking Test, Level X & Level Z-Manual* (3rd ed.). Midwest.

Govier, T. (1987). *Problems in argument analysis and evaluation*. Foris Publications.

Halpern, D. F. (2003). The "how" and "why" of critical thinking assessment. In D. Fasko (Ed.), *Critical thinking and reasoning. Current research, theory, and practice.* (pp. 355–366). Hampton Press.

Halpern, D. F. (2012). *Halpern Critical Thinking Assessment.* Schuhfried, Vienna Test System.

Halpern, D. F. & Dunn, D. S. (2021). *Thought and knowledge: An introduction to critical thinking* (6th ed). Taylor & Francis.

Heard, J., Scoular, C., Duckworth, D., Ramalingam, D., & Teo, I. (2020). Critical thinking: Skill development framework. *Australian Council for Educational Research*. https://research.acer.edu.au/ar_misc/41

Hu, L. T., & Bentler, P. M. (1999). Cutoff criteria for fit indices in covariance structure analysis: Conventional criteria versus new alternatives. *Structural Equation Modeling: A Multidisciplinary Journal, 6*(1), 1–55. https://dx.doi.org/10.1080/10705519909540118

Kretzschmar, A., Neubert, J. C., Wusternberg, S., & Greiff, S. (2016). Construct validity of complex problem-solving: A comprehensive view on different facts of intelligence and school grades. *Intelligence, 54*, 55-69. http:// dx.doi.org/ 10.1016/ j.inell.2015.11,004

MacCallum, R. C., Browne, M. W., & Sugawara, H. M. (1996). Power analysis and determination of sample size for covariance structure modeling. *Psychological Methods, 1*(2), 130–149. https://doi.org/10.1037/1082-989X.1.2.130

Muthén, L. K., & Muthén, B. O. (2019). *Mplus statistical modeling software: Release 8.6*. Muthén & Muthén.

Nieto, A. M., & Saiz, C. (2008). Evaluation of Halpern's "Structural Component" for improving Critical Thinking. *Spanish Journal of Psychology, 11*(1), 266–274.

Nieto, A. M., Saiz, C., & Orgaz, B. (2009). Análisis de las propiedades psicométricas de la versión española del HCTAES-Test de Halpern para la evaluación del pensamiento crítico mediante situaciones cotidianas. *Revista Electrónica de Metodología Aplicada, 14*(1), 1–15.

Rivas, S. F., Morales, P., & Saiz, C. (2014). Propiedades psicométricas de la adaptación peruana de la prueba de pensamiento crítico PENCRISAL. *Avaliação Psicológica, 13*(2), 257–268.

Rivas, S. F., & Saiz, C. (2012). Validación y propiedades psicométricas de la prueba de pensamiento crítico PENCRISAL. *Revista Electrónica de Metodología Aplicada*, *17*(1), 18–34.

Rivas, S. F., Saiz, C., & Almeida, L. S. (2020). Pensamiento crítico y el reto de su evaluación. *Educação: Teoria e Prática. 30* (63) https://doi.org/10.18675/1981-8106.v30.n.63.s14706.

Saiz, C. (2020). *Pensamiento crítico y eficacia (2ª ed.)*. Pirámide.

Saiz, C., & Rivas, S. F. (2008). Evaluación del pensamiento crítico: una propuesta para diferenciar formas de pensar. *Ergo, Nueva Época, 22-23*, 25–66. https://doi.org/10.18675/1981-8106.v30.n.63.s14706

Saiz, C., Rivas, S. F., & Almeida, L. S. (2020). Los cambios necesarios en la enseñanza superior que seguro mejorarían la calidad de la educación**.** *E-Psi. Revista Eletrónica de Psicologia, Educação e Saúde, 9*(1), 9–26. http://www.revistaepsi.com

Uribe, E. O. L., Uribe, E. D. S., & Vargas, M. D. P. (2017). Pensamiento crítico y su importancia en la educación: algunas reflexiones. *Rastros Rostros, 19*(34). https://doi.org/10.16925/ra.v19i34.2144

Wechsler, S., Saiz, C., Rivas, S. F., Vendramini, C., Almeida, L. F., Mundim, C. M., & Franco, A. (2018). Creative and critical thinking: Independent or overlapping components? *Thinking Skills and Creativity*, *27*, 114–122. https://doi.org/10.1016/j.tsc.2017.12.003

Recebido em: 12/06/2021
Reformulado em: 16/08/2021
Aprovado em: 16/08/2021

Sobre los autores:

Carlos Saiz és Doctor en Psicología, lleva dos décadas enseñando e investigando en pensamiento crítico. Ha publicado numerosos artículos en revistas internacionales e impartido cursos y conferencias en diferentes países. Desde hace unos años, es coordinador del Grupo de Investigación Reconocido (GIR), de la Universidad de Salamanca: Pensamiento Crítico y Psicología Positiva, que tiene dos prioridades: la instrucción y la evaluación en pensamiento crítico.
ORCID: https://orcid.org/0000-0002-5243-958X
E-mail: csaiz@usal.es

Leandro S. Almeida és Doutor em Psicologia (Psicologia da Educação), pela Universidade do Porto (Portugal). Professor de cognição e aprendizagem, e etodología da investigação. Membro do Centro de Investigação em Educação (CIEd – Universidade do Minho), pesquisando nos últimos anos sobre adaptação e sucesso académico dos estudantes do Ensino Superior. Autor ou coautor de testes e questionários.
ORCID: https://orcid.org/0000-0002-0651-7014.
E-mail: silviaferivas@usal.es

Silvia F. Rivas és Doctora en Psicologí, lleva diecinueve años impartiendo clases sobre pensamiento crítico en la Universidad dando cursos y conferencias en otras universidades, de diferentes países. Ha publicado artículos sobre instrucción y evaluación del pensamiento crítico. Actualmente, sigue desarrollando las herramientas de instrucción y evaluación del pensamiento crítico, en la Universidad de Salamanca, dentro del GIR-USAL.
ORCID: https://orcid.org/0000-0002-9790-035X
E-mail: silviaferivas@usal.es

Dirección postal:

Prof. Carlos Saiz
Universidad de Salamanca
Facultad de Psicología
Avda. de la Merced, 109-131
37005 Salamanca. España