ENSEÑANZA DE LAS CIENCIAS, 38-3 (2020), 143-159
Investigaciones didácticas
https://doi.org/10.5565/rev/ensciencias.2854
ISSN (impreso): 0212-4521 / ISSN (digital): 2174-6486

Revisión sistemática de instrumentos de actitudes hacia la ciencia (2004-2016)

Systematic review of attitude toward science instruments (2004-2016)

Radu Bogdan Toma
Departamento de Didácticas Específicas. Facultad de Educación. Universidad de Burgos. Burgos, España.
rbtoma@ubu.es

RESUMEN • Se realiza una revisión sistemática de las propiedades psicométricas que poseen 15 instrumentos de actitudes hacia la ciencia publicados entre los años 2004 y 2016, identificados en las bases de datos *Web of Science Core Collection* y *Science Direct* siguiendo los criterios PRISMA. En conjunto, los resultados revelan que la calidad de la mayoría de los instrumentos actuales sigue siendo preocupante dada la ausencia casi generalizada de adopción de un marco teórico para el desarrollo de los ítems y debido a un deficiente e inadecuado análisis de sus propiedades psicométricas. Estos hallazgos cuestionan la validez y fiabilidad de los instrumentos analizados y ponen de manifiesto la necesidad de adoptar diseños metodológicos más rigurosos para el desarrollo de instrumentos de actitudes hacia la ciencia.

PALABRAS CLAVE: Actitudes hacia la ciencia; Desarrollo y validación; Instrumentos; Propiedades psicométricas; Revisión sistemática.

ABSTRACT • A systematic review of the psychometric properties of 15 attitudes towards science instruments published between 2004 and 2016, identified in the *Web of Science Core Collection* and *Science Direct* databases following the PRISMA criteria, is reported. Overall, the quality of most examined instruments continues to be worrying due to an almost generalized absence of adoption of a theoretical framework for item development and due to a scarce and inadequate analysis of their psychometric properties. These findings calls into question the validity and reliability of the tools examined and highlight the need to adopt more rigorous methodological designs for the development of attitude toward science measurement strategies.

KEYWORDS: Attitudes towards science; Development and validation; Tools; Psychometric properties; Systematic review.

Recepción: diciembre 2018 • Aceptación: junio 2019 • Publicación: noviembre 2020

Toma, R. B. (2020).Revisión sistemática de instrumentos de actitudes hacia la ciencia (2004-2016). *Enseñanza de las Ciencias, 38*(3), 143-159.
https://doi.org/10.5565/rev/ensciencias.2854

INTRODUCCIÓN

En los últimos años, nuevos enfoques educativos para reducir el desinterés hacia la ciencia han cobrado un fuerte impulso y, actualmente, se están promoviendo en gran medida iniciativas enfocadas al fomento de los conocimientos y de las competencias científicas necesarias que permitan a los estudiantes una participación activa y responsable en la sociedad (EC, 2015). Así, uno de los principales objetivos es el desarrollo de actitudes favorables hacia la ciencia (Osborne, Simon y Collins, 2003), con propuestas como *Horizonte 2020*, promovida por la Comisión Europea, que subrayan la necesidad de hacer atractivas las ciencias para todos los jóvenes (EC, 2016). Por lo tanto, se están impulsando medidas educativas para aumentar la exposición de los estudiantes a cuestiones científicas, especialmente desde grados elementales del sistema educativo (NRC, 2011; EC, 2016), siendo numerosas las políticas educativas consistentes con estas recomendaciones (Bybee, 2013; NRC, 2012, 2013). Este énfasis en desarrollar prácticas educativas dirigidas a despertar el interés y el desarrollo de actitudes favorables hacia la ciencia ha sido acompañado por el desarrollo de nuevos instrumentos de autoinforme para la evaluación de estas reformas.

Dados los resultados reportados por Blalock, Lichtenstein, Owen, Pruski, Marshall y Toepperwein (2008) en relación con la falta de evidencias psicométricas de los instrumentos de actitudes hacia la ciencia publicados hasta el año 2005, resulta necesario evaluar si los instrumentos diseñados en los últimos años poseen propiedades psicométricas más robustas que los publicados con anterioridad. Aunque algunas investigaciones han abordado esta cuestión recientemente, aún no se ha realizado un análisis exhaustivo de las propiedades psicométricas de los instrumentos desarrollados y publicados después de la revisión de Blalock et al. (2008). Por ejemplo, Aydeniz y Kotowski (2014) informaron sobre algunas limitaciones conceptuales de apenas cinco instrumentos de los comúnmente utilizados para medir las actitudes hacia la ciencia y Potvin y Hasni (2014) proporcionaron información descriptiva sobre algunos de los instrumentos sin examinar la calidad de estos sobre la base de las propiedades psicométricas.

Por esta razón, el propósito de este estudio consiste en actualizar el trabajo de Blalock et al. (2008) identificando y analizando los instrumentos de actitudes hacia la ciencia publicados en los últimos años para determinar en qué medida los nuevos esfuerzos que se están promoviendo para la mejora de la educación científica están siendo complementados con el desarrollo de instrumentos de autoinforme válidos y fiables.

MÉTODO

En la realización de este trabajo se ha seguido el procedimiento descrito por Bennett, Lubben, Hogarth y Campbell (2005) para la realización de revisiones sistemáticas y, la declaración PRISMA (Liberati et al., 2009) para el reporte transparente de los resultados derivados de revisiones sistemáticas.

Las revisiones sistemáticas difieren de las narrativas en que son menos propensas a sesgos de investigación, por lo que se introduce una mayor objetividad al explicar rigurosa y explícitamente el proceso de selección de los estudios y el método utilizado para su revisión y evaluación. Por lo tanto, las revisiones sistemáticas «(...) incluyen una fase de codificación más detallada en la que se identifican las características clave de los estudios con el fin de proporcionar una visión general del trabajo en el área bajo consideración» (ibíd., p. 390 [traducción propia]). Por este motivo, las revisiones sistemáticas tienden a ser más transparentes y reproducibles que las revisiones tradicionales. En consecuencia, a continuación, se detallan las principales fases de este estudio de acuerdo con el procedimiento introducido por Bennett et al. (2005).

Criterio de elegibilidad

Debido a la multiplicidad de definiciones existentes en la literatura relacionada con las actitudes hacia la ciencia, es necesario proveer primero una definición operativa del constructo objeto de estudio. Si bien en el contexto español existen autores españoles muy prolíficos en esta línea de investigación, como por ejemplo Vázquez y Manassero (1995), que han propuesto una taxonomía de las actitudes hacia la ciencia, este estudio pretende ser un *continuum* del de Blalock et al. (2008), por lo que se ha adoptado la misma conceptualización de actitudes hacia la ciencia que estos autores que, además, se encuentra en consonancia con la empleada en contextos iberoamericanos. Así, siguiendo a Gardner (1975), se define las actitudes hacia la ciencia como las respuestas emocionales hacia lo científico, es decir, hacia la ciencia en general y hacia las disciplinas, carreras o asignaturas científicas en particular. Por tanto, se ha establecido que serán elegibles aquellos estudios cuya conceptualización de las actitudes hacia la ciencia sea coherente con esta definición. Respuestas emocionales como interés –«Espero con ganas las clases de ciencias» (Wang y Berlin, 2010, p. 2423 [traducción propia])–, apreciación –«La ciencia puede ayudar a hacer del mundo un lugar mejor» (Navarro et al., 2015, p. 1465)– o disfrute –«Creo que aprender ciencias es divertido» (Zhang y Campbell, p. 601 [traducción propia])– son ejemplos que reflejan actitudes hacia la ciencia.

Búsqueda bibliográfica

Se identificaron artículos potencialmente relevantes para el objetivo de este estudio en las bases de datos *Web of Science Core Collection* y *Science Direct*, empleando una estrategia de búsqueda consistente con el criterio de elegibilidad. En la base de datos de *Web of Science Core Collection* se utilizó una combinación de los términos «actitud* hacia la* ciencia*» como palabras clave en los títulos e «instrumento* OR escala* OR medida*» como palabras clave en el tema, tanto en castellano como en inglés;[1] se restringió la búsqueda por categoría de investigación (educación e investigación educativa), tipo de documento (artículos), idioma (inglés y español) y año de publicación (2004-2016). En la base de datos *Science Direct*, estos términos se emplearon como palabras clave en todos los campos y se restringió la búsqueda por categoría de investigación (ciencias sociales), tipo de documento (artículos) y año de publicación (2004-2016). Debido a que el instrumento más reciente revisado por Blalock et al. (2008) databa del año 2003, se estableció el año 2004 como año inicial y, dado que la búsqueda en las bases de datos mencionadas se realizó durante la primera semana de agosto de 2017, se estableció el año 2016 como el último año de publicación de artículos.

Tras seleccionar los artículos relevantes de las bases de datos, se utilizó un enfoque de ascendencia, que consiste en examinar la lista de referencias de los artículos seleccionados en busca de artículos potencialmente relevantes que no han sido identificados empleando la estrategia de búsqueda descrita anteriormente.

Criterios de inclusión y exclusión

Se formularon los siguientes criterios de inclusión y exclusión, atendiendo al criterio de elegibilidad en este estudio: (i) estudios de desarrollo y validación de instrumentos, (ii) centrados en el constructo de actitudes hacia la ciencia, (iii) que emplean la teoría clásica de test (*Classical Test Theory*) para la validación del instrumento, y (iv) cuyos instrumentos sean de naturaleza cuantitativa. Se excluyeron los estudios que emplean instrumentos de actitudes hacia la ciencia y que, sin embargo, no están específicamente enfo-

1. En inglés, se han empleado los siguientes términos: «attitude* to* science*» e «instrument* OR scale* OR measure*».

cados en su desarrollo y validación, los que no cumplen con el criterio de elegibilidad en el que se define el constructo de actitud hacia la ciencia adoptado en este estudio, los que no utilizan la teoría clásica de pruebas para la validación del instrumento (es decir, los que utilizan el análisis Rasch) y los que emplean métodos cualitativos de recopilación de datos (por ejemplo, entrevistas o dibujos de los estudiantes).

Rúbrica de evaluación

Se empleó la rúbrica desarrollada por Blalock et al. (2008) para evaluar las propiedades psicométricas de los instrumentos incluidos en este estudio, lo que podría permitir la comparación entre los instrumentos incluidos en este trabajo y los analizados en el estudio de Blalock et al. (2008). No obstante, han sido necesarias ligeras adaptaciones. La rúbrica original constaba de cinco secciones: (i) marco teórico, (ii) fiabilidad, (iii) validez, (iv) dimensionalidad y (v) uso, con una evaluación total posible de 28 puntos. La última sección pretendía evaluar hasta qué punto los instrumentos desarrollados habían sido empleados en estudios posteriores al de su desarrollo y validación. Sin embargo, una revisión inicial de los artículos incluidos en este trabajo reveló que aproximadamente la mitad de ellos fueron publicados entre los años 2014 y 2016. Por lo tanto, esta sección de la rúbrica fue descartada en este estudio para asegurar una igualdad de condiciones y no perjudicar los instrumentos publicados en años posteriores, que cuentan con menos probabilidad de haber sido utilizados en otros estudios. Este cambio redujo la puntuación total posible de 28 a 27 puntos (tabla 1).

Tabla 1.
Rúbrica de evaluación

Aspecto	*Puntuación*			
Marco teórico (0-3)				
Antecedentes teóricos para el desarrollo del instrumento	0 (No)	3 (Sí)		
Fiabilidad (0-9)				
Consistencia interna	0 (no reportado o $\alpha < 0,60$)	1 ($\alpha = 0,61$-$0,70$)	2 ($\alpha = 0,71$-$0,80$)	3 ($\alpha > 0,80$)
Test-retest	0 (no reportado o $r < 0,60$)	1 ($r = 0,61$-$0,70$)	2 ($r = 0,71$-$0,80$)	3 ($r > 0,80$)
Error estándar de medición	0 (No)	3 (Sí)		
Validez (0-9)				
Contenido, constructo, convergente, concurrente, discriminante, discriminativa, predictiva	0 (no reportado)	3 (1-2 evidencias)	6 (3-4 evidencias)	9 (>4 evidencias)
Dimensionalidad (0-6)				
En instrumentos unidimensionales, se ha empleado la puntuación global. En instrumentos multidimensionales, se ha empleado la puntuación de cada dimensión.	0 (No)	3 (Sí)		
En instrumentos sometidos a análisis factorial, ¿las subescalas reflejan los factores adecuados?	0 (No)	3 (Sí)		
	Puntuación total 0-27			

Adicionalmente, dados los avances en las definiciones y tipos de propiedades psicométricas surgidos desde el desarrollo de la rúbrica hace ya más de una década, especialmente en relación con las pruebas de validez de los instrumentos, se ha empleado la taxonomía de Polit (2015) y Polit y Yang (2016) para actualizar la terminología referente a la validez de los instrumentos, sustituyéndose tres términos empleados por Blalock et al. (2008) por los propios de la taxonomía revisada y añadiéndose dos criterios de validez adicionales no contemplados en la rúbrica original. Estos cambios, así como la explicación de cada prueba de fiabilidad y validez, se recogen en la tabla 2.

Tabla 2.
Taxonomía revisada y explicación de cada propiedad psicométrica

Blalock et al. (2008)	*Taxonomía revisada*[a]	*Explicación*[a]
Fiabilidad		
Consistencia interna	Consistencia interna	Refiere al grado en que los ítems miden el mismo constructo y si los ítems seleccionados están interrelacionados entre sí, utilizando α de Cronbach
Test-retest	Test-retest	Evalúa la estabilidad y reproducibilidad de los resultados, partiendo del supuesto de que el rasgo objeto de estudio no ha cambiado en los sujetos
Error estándar de medición	Error estándar de medición	Examina en qué medida los resultados del instrumento están libres de error de medición.
Validez		
Contenido	Contenido	Refiere a la idoneidad y relevancia de los ítems para representar la naturaleza y dimensionalidad del constructo objeto de estudio.
Congruente	Concurrente	Comprueba si los resultados son coherentes en comparación con otro instrumento de referencia que mide el mismo constructo.
Discriminante	Discriminante	Prueba las hipótesis de que el instrumento mide el constructo objeto de estudio y no un constructo diferente al previsto.
Grupos de contraste	Discriminativa	Examina el grado en que el instrumento discrimina entre grupos que teoréticamente difieren con respecto al constructo objeto de estudio.
Análisis factorial	Constructo	Valora si el instrumento capta la dimensionalidad hipotética del constructo objeto de estudio, utilizando análisis factorial.
	Convergente	Evalúa la correlación entre las puntuaciones del constructo objeto de estudio y las de un constructo de convergencia conceptual.
	Predictiva	Comprueba si el instrumento predice aquellos constructos que teoréticamente debería predecir.

[a]La taxonomía revisada y su explicación se basa en la taxonomía propuesta por Polit (2015) y Polit y Yang (2016).

Fiabilidad en el análisis de los instrumentos

Para asegurar que cada instrumento fuera evaluado de manera correcta y consistente, se efectuó una medición formal de la fiabilidad intraevaluador en la aplicación de la rúbrica. El autor de este estudio aplicó dos veces la rúbrica de evaluación a cada instrumento con un período de diferencia de 6-8 semanas entre la primera y la segunda aplicación de la rúbrica y calculó la correlación entre los dos conjuntos de datos para examinar si existen discrepancias entre las dos aplicaciones.

El índice de correlación de Pearson reportó una relación significativa ($p < 0{,}001$) entre los dos conjuntos de evaluaciones de cada sección de la rúbrica. Los datos de la primera y segunda aplicación de la rúbrica estuvieron perfectamente correlacionados ($r = 1$) en las secciones «Marco teórico», «Test re-test», «Error estándar de medición» y «Dimensionalidad» y, fuertemente correlacionados en las sec-

ciones «Consistencia interna» (r = 0,98) y «Validez» (r = 0,93). Los instrumentos en los que no se ha identificado una correlación perfecta entre la primera y la segunda evaluación fueron examinados por tercera vez para resolver las discrepancias existentes. Las discrepancias encontradas consistieron en asignar una consistencia interna ligeramente mayor a una dimensión de un instrumento y en no identificar una validez de contenido en un instrumento. Cabe destacar que estas discrepancias no afectaron a la puntuación total asignada a los instrumentos, por lo que los resultados de este estudio poseen un alto nivel de fiabilidad.

RESULTADOS

Identificación y selección de estudios

En la figura 1 se recoge el proceso de identificación y selección de los estudios atendiendo a los criterios PRISMA. En la primera fase, llamada *identificación*, 104 y 54 referencias fueron recuperadas de las bases de datos *Web of Science* y *Science Direct*, respectivamente. En la segunda fase, denominada *cribado*, se aplicaron los filtros de búsqueda, reteniéndose un total de 65 artículos. En la tercera fase, *elegibilidad*, se excluyeron un total de 54 artículos tras la lectura completa de estos y la aplicación de los criterios de inclusión. Finalmente, 14 artículos fueron retenidos para su análisis en profundidad en la fase de *inclusión*, once de los cuales fueron derivados de las bases de datos consultadas y tres mediante el enfoque de ascendencia.

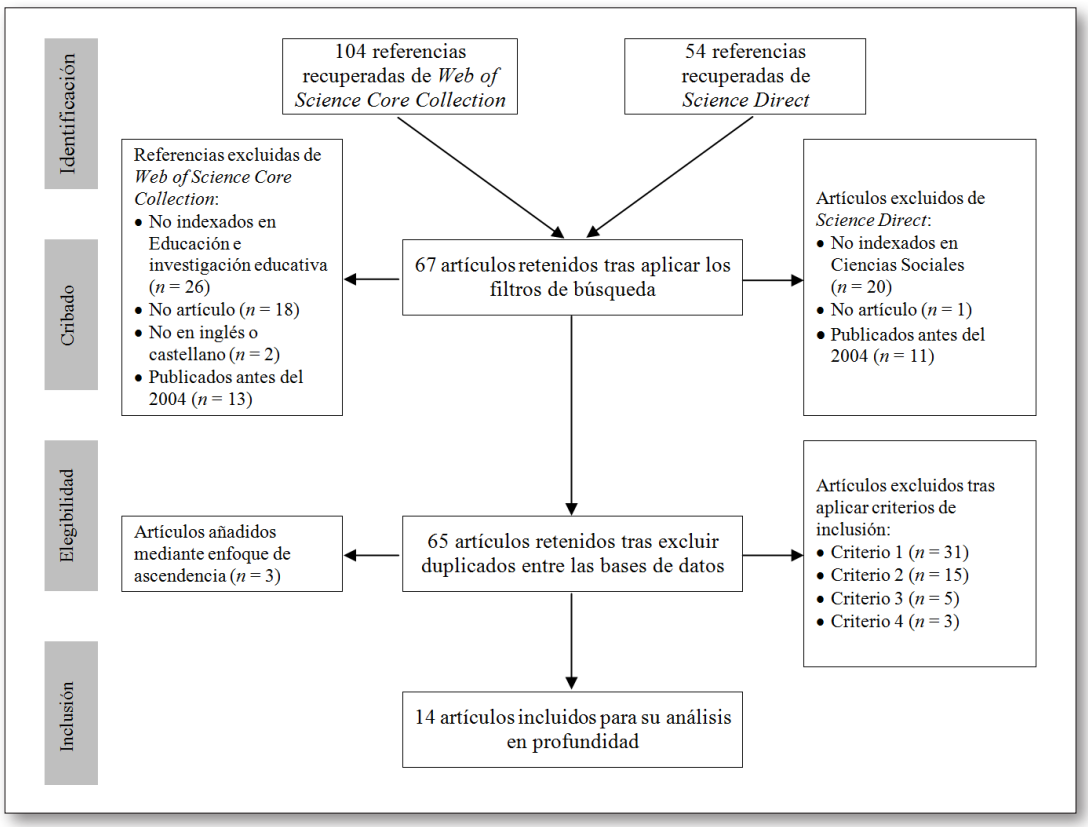

Fig. 1. Identificación y selección de estudios según PRISMA

Descripción general de los instrumentos

La tabla 3 recoge la síntesis y evaluación de los instrumentos incluidos en este estudio. Los 14 artículos seleccionados ofrecen información sobre el desarrollo y la validación de un total de 15 instrumentos. Cinco de estos instrumentos fueron validados empleando muestras de la etapa primaria (hasta 10-11 años, 5.º curso EPO), diez de la etapa media (hasta 13-14 años, 6.º EPO - 2.º ESO), tres en la etapa de secundaria (hasta 17-18 años, 3.º ESO - 2.º Bachillerato), dos para estudiantes universitarios y uno para profesorado.[2] Un total de 12 instrumentos fueron desarrollados con muestras de habla inglesa. De los restantes, el instrumento de Abd-El-Khalick, Summers, Said, Wang y Culbertson (2015) fue validado con muestras de habla árabe e inglesa y el de Navarro, Förster, González y González-Pose (2016) con muestras de habla española. No se pudo identificar el idioma objeto de estudio del instrumento desarrollado con estudiantes taiwaneses por Wang y Berlin (2010) pues los autores no indicaron en qué idioma fue administrado.

En relación con la estructura de los instrumentos, tan solo Wang y Berlin (2010) desarrollaron uno de naturaleza unidimensional, siendo los restantes multidimensionales. Por otro lado, 12 instrumentos emplearon ítems de tipo Likert con un formato de 5 opciones de respuesta, uno utilizó ítems de diferencial semántico y otro una combinación de ambos tipos. La longitud de los instrumentos varió considerablemente, siendo el más corto el desarrollado por Kennedy, Quinn y Taylor (2016), compuesto por apenas diez ítems y, el más largo la adaptación al español del TOSRA de Fraser (1981), realizada por Navarro et al. (2016), que consta de un total de 70 ítems.

Cabe destacar que algunos instrumentos ampliamente utilizados en el contexto español han sido descartados atendiendo al primer criterio de inclusión. Así, por ejemplo, los instrumentos PANA (de Pro y Pérez, 2014) y el ROSE, empleado principalmente por Vázquez y Manassero (2009a), Fernández-Cézar, Pinto-Solano y Muñoz-Hernández (2018) y Marbá-Tallada y Márquez (2010), son empleados en estudios cuyo objetivo principal no es el desarrollo y la validación del propio instrumento. Por otro lado, aunque Vázquez y Manassero (2008, 2009*b*) han analizado la estructura factorial del instrumento ROSE, estos artículos tampoco están enfocados a la validación de este instrumento.

Tabla 3.
Resumen y evaluación de los instrumentos analizados

Autores	*Muestra*[a]	*Ítems*	*Tipo de ítems*	*Marco teórico (0-3)*	*Fiabilidad (0-9)*[b]	*Validez (0-9)*	*Dimensionalidad (0-6)*	*Puntuación total (0-27)*
Abd-El-Khalick, Summers, Said, Wang y Culbertson (2015)	A B C	32	Likert	Sí (3)	,61-,87[c] (1)	Contenido Constructo (3)	Sí (3) Sí (3)	13
Dijkstra y Goedhart (2012)	B	38	Likert	No (0)	α = ,71-,87 (2)	Contenido Constructo Convergente Discriminativa (6)	Sí (3) Sí (3)	14

2. Algunos instrumentos son aptos para varias etapas educativas, tal y como se indica en la tabla 3.

Autores	*Muestra*[a]	*Ítems*	*Tipo de ítems*	*Marco teórico (0-3)*	*Fiabilidad (0-9)*[b]	*Validez (0-9)*	*Dimensionalidad (0-6)*	*Puntuación total (0-27)*
Guzey, Harwell y Moore (2014)	A	28	Likert	No (0)	α = ,77-,87 (2)	Contenido Constructo Discriminativa (6)	Sí (3) Sí (3)	14
Hillman, Zeeman, Tilburg y List (2016)	A B C	40	Likert	No (0)	α = ,43-,91 (0)	Contenido (3)	-	3
Kennedy, Quinn y Taylor (2016)	B	10	Likert Diferencial semántico	No (0)	α = ,82-,98 (3)	Contenido Constructo (3)	Sí (3) Sí (3)	12
Kind, Jones y Barmby (2007)	B	45	Likert	No (0)	α = ,72-,94 (2)	Constructo Convergente (3)	Sí (3) Sí (3)	11
Mahoney (2010)	C	23	Likert	No (0)	α = ,76-,96 (2)	Contenido Discriminativa (3)	-	5
Navarro, Förster, González y González-Pose (2016)s	B	70	Likert	Sí (3)	α = ,63-,91 (1)	Constructo Convergente Discriminante (6)	Sí (3) Sí (3)	16
Owen et al. (2008)	B	22	Likert	No (0)	α = ,75-,78 (2)	Constructo Discriminativa (3)	Sí (3) Sí (3)	11
Puvirajah, Verma, Li y Martin-Hansen (2015)	B	22	Likert	No (0)	α = ,83-,90 (3)	Constructo (3)	Sí (3) Sí (3)	12
Tyler-Wood, Knezek y Christensen (2010)	B	12	Likert	No (0)	α = ,78-,94 (2)	Contenido Constructo Convergente (6)	Sí (3) Sí (3)	14
Tyler-Wood, Knezek y Christensen (2010)	B D E	25	Diferencial semántico	No (0)	α = ,84-,93 (3)	Contenido Constructo Discriminativa (6)	Sí (3) Sí (3)	15
Villafañe y Lewis (2016)	D	24	Likert	Sí (3)	α = ,77-,87 (2)	Constructo Discriminativa Predictiva (6)	Sí (3) Sí (3)	17
Wang y Berlin (2010)	A	30	Likert	No (0)	α = ,93 (3)	Contenido Constructo (3)	Sí (3) No (0)	9
Zhang y Campbell (2011)	A	28	Likert	Sí (3)	α = ,65-,88 (1)	Constructo Convergente (3)	Sí (3) Sí (3)	13

[a]Los cursos equivalentes a España son: A (hasta 5.º EPO); B (6.º EPO hasta 2.º ESO); C (3.º ESO hasta 2.º Bachillerato); D (universitarios); E (profesorado). [b]La puntuación ha sido otorgada atendiendo al valor más bajo de fiabilidad. [c]Los autores han empleado estadísticos de fiabilidad a partir de análisis factorial confirmatorio en lugar de α de Cronbach.

Calidad psicométrica de los instrumentos

En conjunto, la puntuación final de los instrumentos osciló entre 3 y 17 puntos (figura 2). Más de la mitad de los instrumentos (9 de 15, 60 %) obtuvieron una puntuación inferior a la mitad de la puntuación total de la rúbrica. El instrumento desarrollado por Villafañe y Lewis (2016) obtuvo la puntuación más alta (17 de un total de 27 puntos posibles) y el instrumento de Hillman et al. (2016) fue el que obtuvo menor puntuación, con apenas 3 de los potenciales 27 puntos.

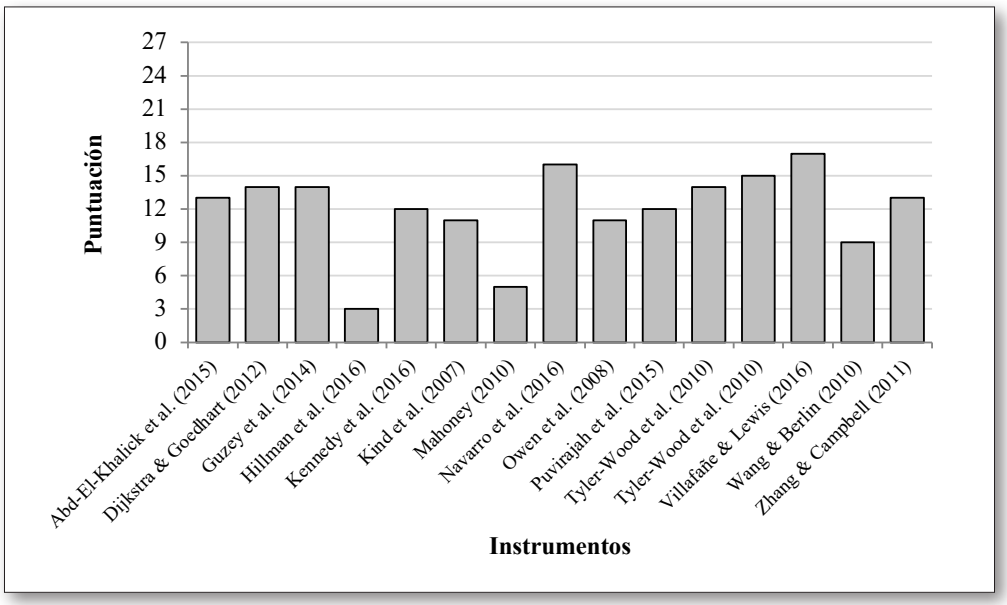

Fig. 2. Puntuación total de los instrumentos analizados.

Atendiendo a los resultados según las secciones de la rúbrica empleada, se observó que 11 de los 15 instrumentos analizados fueron desarrollados sin un marco teórico de actitudes hacia la ciencia (tabla 3). De los estudios que sí habían empleado un marco teórico, Abd-El-Khalick et al. (2015) desarrolló un instrumento basado en la Teoría de la Conducta Planeada de Ajzen (1991); Navarro et al. (2016) y Villafañe y Lewis (2016) emplearon la clasificación de actitudes planteada por Klopfer (1971) y posteriormente popularizada por Fraser (1981) y su instrumento TOSRA; finalmente, Zhang y Campbell (2011) emplearon la conceptualización tripartita según la cual las actitudes se clasifican en componentes afectivos, cognitivos y conductuales (Breckler, 1984; Eagly y Chaiken, 1998).

En relación con la fiabilidad de los instrumentos, ningún estudio reportó evidencias psicométricas de fiabilidad temporal (test-retest) ni información acerca del error estándar de medición de los instrumentos desarrollados. Todos los estudios menos el de Abd-El-Khalick et al. (2015), que empleó índices de fiabilidad a partir de análisis factorial confirmatorio, reportaron el índice de α de Cronbach como única evidencia de fiabilidad. No obstante, en la figura 3 se observa que cuatro instrumentos poseen una consistencia interna inferior a la mínima aceptada para estudios exploratorios, según Nunnally (1978), y menos de una cuarta parte de los instrumentos poseen niveles de fiabilidad robustos.

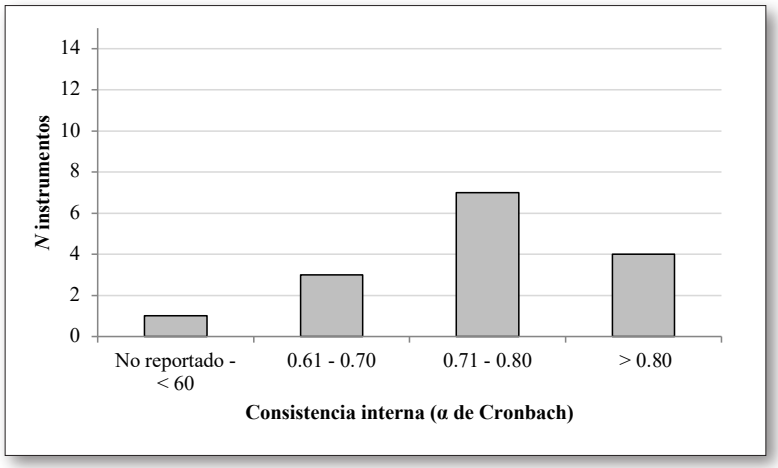

Fig. 3. Consistencia interna de los instrumentos analizados.

En cuanto a la validez de los instrumentos, más de la mitad (9 de 15, 60 %) apenas reportaron una o dos evidencias psicométricas y los seis instrumentos restantes analizados proporcionaron tres o cuatro evidencias de validez. Cabe destacar que ningún instrumento fue sometido a más de cuatro pruebas de validez psicométrica.

De las posibles pruebas psicométricas disponibles para analizar la validez de un instrumento, la validez de constructo fue la más empleada, seguida por la validez de contenido (figura 4). Ningún instrumento fue sometido a pruebas psicométricas de validez concurrente y apenas un instrumento proporcionó información sobre la validez discriminante y predictiva.

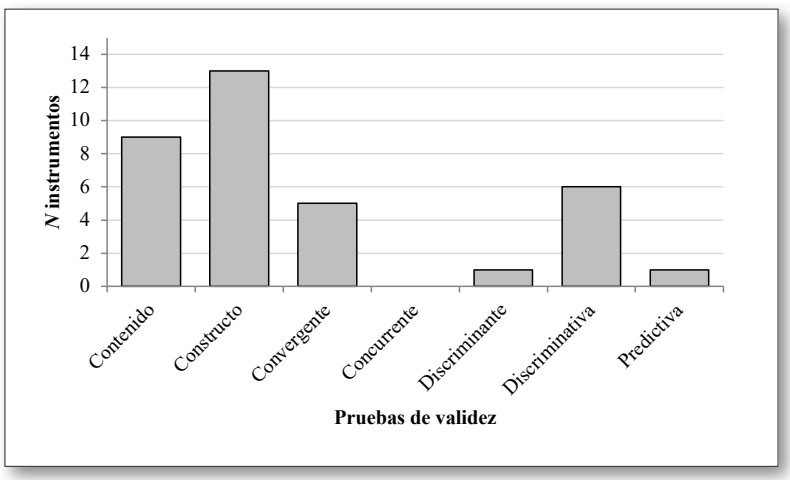

Fig. 4. Frecuencia de pruebas de validez reportadas en los instrumentos analizados.

DISCUSIÓN

Los resultados de esta revisión indican que las críticas realizadas por Blalock et al. (2008) hace más de una década siguen estando vigentes para los estudios actuales enfocados al desarrollo y a la validación de instrumentos de actitudes hacia la ciencia. Además, lo que resulta aún más preocupante es descubrir que las deficiencias en los instrumentos analizados por Munby (1983) hace casi cuarenta años han continuado en esta línea de investigación hasta la actualidad. Así, aunque el uso de un marco teórico resulta

imprescindible para el desarrollo de instrumentos, los estudios revisados se caracterizan por la ausencia de marco teórico para la conceptualización de qué constituyen las actitudes hacia la ciencia. Por lo tanto, parece ser que los autores de estos instrumentos se han limitado meramente a crear o adaptar ítems a partir de cuestionarios existentes (en la mayoría de los casos también con una pobre conceptualización teórica de las actitudes) y, a emplear dimensiones actitudinales de forma aleatoria y sin una base teórica coherente. Ello lleva a que se incluya bajo el mismo *paraguas* constructos tan diversos –y que cuentan con marcos teóricos propios diferenciados– como creencias y valores (Hilman et al., 2016), autoeficacia y motivaciones (Puvirajah et al., 2015) o habilidad percibida, valores y compromiso (Mahoney, 2010). Asimismo, la ausencia de un marco teórico conduce a la inclusión de constructos cuya relevancia en el estudio de las actitudes hacia la ciencia es en el mejor de los casos dudosa. Por ejemplo, Kind et al. (2007) han incluido un constructo que mide las actitudes hacia la escuela en general y un constructo que mide el «interés combinado» hacia la ciencia; Dijkstra y Goedhart (2012) han mezclado constructos tan dispares como actitudes hacia la ciencia escolar, los científicos, la necesidad de abordar el cambio climático y, comportamientos proambientales sin ofrecer justificación alguna, por lo que la inclusión de estos aspectos en un mismo instrumento resulta cuando menos cuestionable.

El segundo aspecto que resulta preocupante es la escasez de estudios que han sometido los instrumentos desarrollados a pruebas psicométricas de validez concurrente y discriminante para determinar en qué medida el instrumento propuesto aborda el constructo que se pretende medir. Como consecuencia del desuso de estas pruebas psicométricas, no es de extrañar la baja fiabilidad general observada en los instrumentos analizados, claro indicio de que los ítems no miden el mismo constructo latente. Por otro lado, la falta de información sobre el error estándar de medición de los instrumentos plantea la cuestión de en qué medida los resultados obtenidos mediante el uso de estos instrumentos son el reflejo de la actitud de los individuos objeto de estudio o es el efecto de un instrumento poco fiable. Asimismo, la ausencia de evidencias de fiabilidad temporal (test-retest) hace que estos instrumentos no sean apropiados para estudios longitudinales o estudios enfocados a la evaluación de intervenciones educativas (diseños pretest-posttest). Dado que la estabilidad y reproducibilidad de los resultados no está garantizada, los investigadores que evalúan intervenciones educativas emplearían estos instrumentos *a ciegas* y sin posibilidad alguna de conocer con certeza en qué medida la intervención educativa ha sido eficaz.

El tercer aspecto preocupante es la proliferación del uso de la validez aparente (en inglés, *face validity*) como una evidencia de validez psicométrica, o el mal uso de algunas pruebas psicométricas. Por ejemplo, Kennedy et al. (2016) y Wang y Berlin (2010), entre otros, han empleado la prueba de validez aparente pese a que es bien sabido que se trata de una práctica poco apropiada (DeVellis, 2003) y que, por consiguiente, los instrumentos deben ser sometidos a una validez de contenido mediante el uso de un panel compuesto por expertos en el constructo objeto de estudio. En cuanto al mal uso de algunas pruebas, por ejemplo, Tytler-Wood et al. (2010) han realizado un análisis factorial con un tamaño muestral inferior a 100 casos. Otro ejemplo de prácticas pocos recomendadas está relacionado con la gran mayoría de los estudios que han realizado análisis factorial exploratorio y que han empleado el método *Little Jiffy* (Kaiser, 1970), práctica que consiste en elegir un método de extracción de componentes principales, una rotación ortogonal Varimax y determinar el número de factores a extraer basado en el criterio Kaiser > 1. Este procedimiento ha sido duramente criticado en la literatura por la falta de robustez en los resultados que proporciona y por extraer más dimensiones de las que realmente subyacen al constructo objeto de estudio, por lo que se trata de un método ampliamente desaconsejado para constructos en los que se espera una alta correlación entre las dimensiones (Gaskin y Happell, 2014; Lloret-Segura, Ferreres-Traves, Hernández-Baeza y Tomás-Marco, 2014).

Por último, todos los instrumentos analizados en este estudio han sido desarrollados empleando ítems que carecen de una base empírica, por lo que han sido formulados siguiendo el criterio del autor

del instrumento o a partir de instrumentos existentes. Si bien se trata de una práctica extendida en el desarrollo y la validación de instrumentos en general, Ryan y Aikenhead (1992) advirtieron que estos ítems suelen reflejar las ideologías de los investigadores en lugar de medir con precisión las actitudes reales de los individuos objeto de estudio. Por ello, estos autores propusieron desarrollar instrumentos basados en las respuestas de los estudiantes durante entrevistas, dotando así a los ítems de una naturaleza empírica, siendo esta recomendación ampliamente ignorada en los instrumentos analizados en este estudio.

IMPLICACIONES

De los resultados de este estudio se derivan varias implicaciones. Debido al panorama desolador de la calidad psicométrica de los instrumentos disponibles para la medición de las actitudes hacia la ciencia, cabría preguntarse por el grado de confianza que se puede depositar en los resultados derivados del uso de estos instrumentos cuya validez y fiabilidad están cuestionadas. Ello requiere, en primera instancia, una reconceptualización del constructo de actitud hacia la ciencia y, posteriormente, el desarrollo de instrumentos que hayan sido sometidos a un riguroso proceso de diseño y validación. Solo de este modo se podría disponer de medidas de autoinforme con propiedades psicométricas más robustas que los instrumentos disponibles actualmente, lo que ayudaría a obtener resultados válidos y fiables que ayuden a respaldar (o en su caso refutar) los supuestos establecidos y consensuados en esta línea de investigación.

Por otro lado, dada la escasez de instrumentos identificados en esta revisión, y en la de Blalock et al. (2008), que hayan sido desarrollados en el contexto español, futuros estudios deberían desarrollar nuevos instrumentos o adaptar y validar instrumentos existentes en otros idiomas para su uso con estudiantes de habla hispana, similar al trabajo de Navarro et al. (2016), revisado en este estudio. De los instrumentos evaluados en este trabajo, quizás los mejores candidatos para este propósito sean el instrumento de Abd-El-Khalick et al. (2015) y el de Zhang y Campbell (2011), que si bien no han sido de los que obtuvieron una mayor puntuación, son instrumentos que han sido desarrollados empleando dos marcos teóricos de amplia trayectoria en el estudio de las actitudes en el ámbito de la psicología social, la teoría de la conducta planeada de Ajzen (1991) en el primero de los casos y la conceptualización tripartita de las actitudes en el segundo.

En este estudio, se han identificado algunos instrumentos que han sido empleados con asiduidad en el contexto español (p. ej., ROSE o PANA). No obstante, estos instrumentos no han sido retenidos para su revisión debido a que han sido utilizados en estudios cuyo objetivo principal no era el análisis de las propiedades psicométricas de estos instrumentos, incumpliendo de este modo el primer criterio de inclusión establecido en este trabajo. Por tanto, parecería necesario que en futuros trabajos se revise la producción iberoamericana para determinar en qué medida el conocimiento actual sobre las actitudes en España y países latinoamericanos ha sido desarrollado sobre la base de instrumentos con una calidad robusta en términos de validez y fiabilidad.

Por último, es necesario desarrollar y consolidar un cuerpo robusto de instrumentos que reporten resultados válidos y fiables que permitan hacer comparaciones entre las actitudes hacia la ciencia identificadas a nivel internacional y local, así como entre la efectividad de las diferentes intervenciones educativas enfocadas a la promoción de actitudes favorables. El uso de un procedimiento e instrumentos similares para la recolección de datos permitiría una mejor comparación de los resultados obtenidos. Por lo tanto, parecería apropiado desarrollar y validar instrumentos para este fin a otros contextos, niveles educativos e idiomas.

LIMITACIONES Y SIGNIFICANCIA DE ESTE ESTUDIO

Los resultados presentados en esta revisión deben interpretarse teniendo en cuenta que solo un investigador participó en el proceso de revisión de todos los instrumentos. Esta limitación se ha tratado de minimizar aplicando la rúbrica de evaluación dos veces a cada uno de los instrumentos y calculando la fiabilidad intraevaluador para obtener resultados fiables. Aunque la principal deficiencia de este método reside en que el evaluador está sujeto a los mismos errores y sesgos, el período de dos meses entre la primera y la segunda evaluación del mismo instrumento ha reducido potencialmente esta deficiencia. Además, el empleo de una rúbrica para la evaluación de los instrumentos ha minimizado el sesgo y los errores que podrían ser introducidos por el investigador, pues el análisis consistió en identificar aquellas pruebas psicométricas que han sido empleadas en cada estudio para validar un instrumento de actitudes hacia la ciencia.

Aunque los resultados de este trabajo deben ser interpretados considerando esta limitación, esta revisión proporciona una evaluación en profundidad de las propiedades psicométricas de los instrumentos de actitudes hacia la ciencia publicados en los últimos años, lo cual puede ser útil por diferentes razones para la investigación en el campo de la didáctica de las ciencias. En primer lugar, en este trabajo se resaltan aquellas prácticas que necesitan ser mejoradas en el desarrollo y la validación de instrumentos. En segundo lugar, este estudio proporciona una guía valiosa que permite a los investigadores examinar las ventajas y desventajas de cada instrumento, resaltando aquellos con mejores y peores propiedades psicométricas. Esto facilita a los autores interesados en el estudio de las actitudes hacia la ciencia la selección de instrumentos de medida de acuerdo con sus necesidades de investigación. En tercer lugar, en esta revisión se ofrece una rúbrica actualizada de los estándares psicométricos modernos, que puede ser utilizada para evaluar los instrumentos de actitudes hacia la ciencia que no han sido incluidos en este trabajo, como, por ejemplo, aquellos empleados en el contexto de la investigación en didáctica de las ciencias desarrollada en España y en Latinoamérica. Asimismo, esta rúbrica podría ser empleada para evaluar la fiabilidad y validez de otros instrumentos de naturaleza cuantitativa que están enfocados a medir otros constructos de relevancia educativa como, por ejemplo, la naturaleza de la ciencia, las motivaciones o la autoeficacia. Por último, este trabajo podría fomentar futuras investigaciones en la línea de las actitudes hacia la ciencia al plantear un debate sobre la necesidad de revisar aquellos supuestos que han sido consensuados a partir de los resultados derivados del uso de instrumentos de limitada validez y fiabilidad.

CONCLUSIONES

Este estudio trata de actualizar el trabajo de Blalock et al. (2008) evaluando las propiedades psicométricas que han sido reportadas en los instrumentos de actitud hacia la ciencia publicados entre los años 2004 y 2016. En conjunto, se puede afirmar que la calidad general de los instrumentos sigue siendo limitada y que se siguen perpetuando las prácticas de desarrollo y validación identificadas como inadecuadas e insuficientes en publicaciones pasadas. Así, la mayoría de los instrumentos de actitudes hacia la ciencia analizados presentan una falta de evidencias psicométricas y una ausencia de marco teórico. La toma de decisiones sobre cómo fomentar actitudes favorables hacia la ciencia derivada de los resultados obtenidos de instrumentos que muestran una falta de validez y fiabilidad y una pobre conceptualización teórica resulta muy arriesgada, lo que subraya la necesidad de adoptar diseños metodológicos más robustos en futuros estudios de validación de instrumentos para la evaluación del relevante constructo de actitudes hacia la ciencia, tratando de superar los aspectos negativos que reducen su validez y fiabilidad, así como la confianza que se puede depositar en los resultados derivados de su uso.

REFERENCIAS

Las referencias marcadas con un asterisco (*) indican estudios incluidos en la revisión sistemática.

Abd-El-Khalick, F., Summers, R., Said, Z., Wang, S. y Culbertson, M. (2015). Development and large-scale validation of an instrument to assess arabic-speaking students' Attitudes toward Science. *International Journal of Science Education*, *37*(16), 2637-2663. *
https://doi.org/10.1080/09500693.2015.1098789

Ajzen, I. (1991). The theory of planned behavior. *Organizational Behavior and Human Decision Processes*, *50*(2), 179-211.
https://doi.org/10.1016/0749-5978(91)90020-T

Aydeniz, M. y Kotowski, M. R. (2014). Conceptual and methodological issues in the measurement of attitudes towards science. *Electronic Journal of Science Education Electronic Journal of Science Education*, *18*(3), 1-24.

Bennett, J., Lubben, F., Hogarth, S. y Campbell, B. (2005). Systematic reviews of research in science education: Rigour or rigidity? *International Journal of Science Education*, *27*(4), 387-406.
https://doi.org/10.1080/0950069042000323719

Blalock, C. L., Lichtenstein, M. J., Owen, S., Pruski, L., Marshall, C. y Toepperwein, M. (2008). In pursuit of validity: A comprehensive review of science attitude instruments 1935-2005. *International Journal of Science Education*, *30*(7), 961-977.
https://doi.org/10.1080/09500690701344578

Breckler, S. J. (1984). Empirical validation of affect, beha ior, and cognition as distinct components of attitude. *Journal of Personality and Social Psychology*, 47, 1191-1205.

Bybee, R. W. (2013). *The case for STEM education: Challenges and opportunities*. Arlington: NSTA press.

de Pro Bueno, A. y Pérez Manzano, A. (2014). Actitudes de los alumnos de Primaria y Secundaria ante la visión dicotómica de la ciencia. *Enseñanza de Las Ciencias*, *32*(3), 111-132.
https://doi.org/10.5565/rev/ensciencias.1015

DeVellis, R. F. (2003). *Scale development: Theory and applications.* Thousand Oaks, Calif.: Sage Publications.

Dijkstra, E. M. y Goedhart, M. J. (2012). Development and validation of the ACSI: Measuring students' science attitudes, pro-environmental behaviour, climate change attitudes and knowledge. *Environmental Education Research*, *18*(6), 733-749.
https://doi.org/10.1080/13504622.2012.662213. *

Eagly, A. y Chaiken, S. (1998). Attitude structure. En D. Gilbert, S. Fiske y G. Lindsay (Eds.), *Handbook of social psychology* (pp. 269-322). Nueva York: McGraw-Hill.

EC (European Commision) (2015). *Science education for responsible citizenship*. Luxemburgo.

EC (European Commision) (2016). *Horizon 2020. Monitoring report 2014*. Luxemburgo.

Fernández-Cézar, R., Pinto-Solano, N. y Muñoz-Hernández, M. (2018). ¿Mejoran los proyectos de divulgación con experimentación la actitud hacia las clases de ciencias? *Revista de Educacion*, *381* (julio-septiembre), 285-307.
https://doi.org/10.4438/1988-592X-RE-2017-381-389

Fraser, B. J. (1981). *Test of science-related attitudes*. Melbourne: Australian Council for Educational Research.

Gaskin, C. J. y Happell, B. (2014). On exploratory factor analysis: A review of recent evidence, an assessment of current practice, and recommendations for future use. *International Journal of Nursing Studies*, *51*, 511-521.
https://doi.org/10.1016/j.ijnurstu.2013.10.005

Guzey, S. S., Harwell, M. y Moore, T. (2014). Development of an instrument to assess attitudes toward science, technology, engineering, and mathematics (STEM). *School Science and Mathematics*, *114*(6), 271-279. *
https://doi.org/10.1111/ssm.12077

Hillman, S. J., Zeeman, S. I., Tilburg, C. E. y List, H. E. (2016). My attitudes toward science (MATS): The development of a multidimensional instrument measuring students' science attitudes. *Learning Environments Research*, *19*, 203-219. *
https://doi.org/10.1007/s10984-016-9205-x

Kaiser, H. (1970). A second generation Little Jiffy. *Psychometrika*, *35*, 401-415.

Kennedy, J., Quinn, F. y Taylor, N. (2016). The school science attitude survey: A new instrument for measuring attitudes towards school science. *International Journal of Research & Method in Education*, *39*(4), 422-445. *
https://doi.org/10.1080/1743727X.2016.1160046

Kind, P., Jones, K. y Barmby, P. (2007). Developing attitudes towards science measures. *International Journal of Science Education*, *29*(7), 871-893. *
https://doi.org/10.1080/09500690600909091

Klopfer, L. E. (1971). Evaluation of learning in science. En B. S. Bloom, J. T. Kastings y G. F. Madaus (Eds.), *Handbook on summative and formative evaluation of student learning* (pp. 559-642). Nueva York: McGraw-Hill.

Liberati, A., Altman, D. G., Tetzlaff, J., Mulrow, C., Gøtzsche, P. C., Ioannidis J. P. A., Clarke, M., Devereaux, P. J., Kleijnen, J. y Moher, D. (2009). The PRISMA statement for reporting systematic reviews and meta-analyses of studies that Evaluate health care interventions: Explanation and elaboration. *Plos Medicine*, *6*(7), e1000100.
https://doi.org/10.1371/journal.pmed.1000100

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

Mahoney, M. P. (2010). Students' attitudes toward STEM: Development of an instrument for high school STEM-based programs. *Journal of Technology Studies*, *36*(1), 24-34. *

Marbà-Tallada, A. y Márquez, C. (2010). ¿Qué opinan los estudiantes de las clases de ciencias? Un estudio transversal de sexto de primaria a cuarto de ESO. *Enseñanza de Las Ciencias*, *28*(1), 19-30.

Munby, H. (1983). *An investigation into the measurement of attitudes in science education*. Columbus, OH: SMEAC Information Reference Center, The Ohio State University.

Navarro, M., Förster, C., González, C. y González-Pose, P. (2016). Attitudes toward science: Measurement and psychometric properties of the test of science-related attitudes for its use in Spanish-speaking classrooms. *International Journal of Science Education*, *38*(9), 1459-1482. *
https://doi.org/10.1080/09500693.2016.1195521

NRC (National Research Council) (2011). *Successful K-12 STEM education: Identifying effective approaches in science, technology, engineering, and mathematics*. *National Academies Press*. Washington, DC: The National Academies Press.

NRC (National Research Council) (2012). *A framework for K-12 science education: Practices, crosscutting concepts, and core ideas*. Washington, DC: National Academy Press.

NRC (National Research Council) (2013). *Next generation science standards: For stated, by states*. Washington, DC: The National Academies Press.

Nunnally, J. C. (1978). *Psychometric theory*. Nueva York: McGraw-Hill.

Osborne, J., Simon, S. y Collins, S. (2003). Attitudes towards science: A review of the literature and its implications. *International Journal of Science Education*, *25*(9), 1049-1079. https://doi.org/10.1080/0950069032000032199

Owen, S. V., Toepperwein, M. A., Marshall, C. E., Lichtenstein, M. J., Blalock, C. L., Liu, Y., Pruski, L. A. y Grimes, K. (2008). Finding pearls: Psychometric reevaluation of the Simpson-Troost attitude questionnaire (STAQ). *Science Education*, *92*(6), 1076-1095. * https://doi.org/10.1002/sce.20296

Polit, D. F. (2015). Assessing measurement in health: Beyond reliability and validity. *International Journal of Nursing Studies*, *52*(11), 1746-1753. https://doi.org/10.1016/j.ijnurstu.2015.07.002

Polit, D. F. y Yang, F. (2016). *Meaurement and the measurement of change: A primer for health professionals*. Filadelfia: Lippincott Williams & Wilkins.

Potvin, P. y Hasni, A. (2014). Interest, motivation and attitude towards science and technology at K-12 levels: A systematic review of 12 years of educational research. *Studies in Science Education*, *50*(August 2016), 85-129. https://doi.org/10.1080/03057267.2014.881626

Puvirajah, A., Verma, G., Li, H. y Martin-Hansen, L. (2015). Influence of a science-focused after-school program on underrepresented high-school students' science attitudes and trajectory: A survey validation study. *International Journal of Science Education Part B-Communication and Public Engagement*, *5*(3), 250-270. * https://doi.org/10.1080/21548455.2014.930210

Ryan, A. G. y Aikenhead, G. S. (1992). Students' preconceptions about the epistemology of science. *Science Education*, *76*, 559-580.

Tyler-Wood, T., Knezek, G. y Christensen, R. (2010). Instruments for assessing interest in STEM content and careers. *Journal of Technology and Teacher Education*, *18*(2), 341-363. *

Vázquez, A. y Manassero, M. A. (1995). Actitudes relacionadas con la Ciencia: una revisión conceptual. *Enseñanza de Las Ciencias*, *13*(3), 337-346.

Vázquez, A. y Manassero, M. A. (2008). La vocación científica y tecnológica de las chicas en secundaria y la educación diferenciada. *Bordón. Revista de Pedagogía*, *60*(3), 149-163.

Vázquez, A. y Manassero, M. A. (2009a). La relevancia de la educación científica: actitudes y valores de los estudiantes relacionados con la Ciencia y la Tecnología. *Enseñanza de Las Ciencias*, *27*(1), 33-48.

Vázquez, A. y Manassero, M. A. (2009b). Factores actitudinales determinantes de la vocación científica y tecnológica en secundaria. *Cultura y Educacion*, *21*(3), 319-330. https://doi.org/10.1174/113564009789052280

Villafañe, S. M. y Lewis, J. E. (2016). Exploring a measure of science attitude for different groups of students enrolled in introductory college chemistry. *Chemistry Education Research and Practice*, *17*, 731-742. * https://doi.org/10.1039/c5rp00185d

Wang, T. y Berlin, D. (2010). Construction and validation of an instrument to measure Taiwanese elementary students' attitudes toward their science class. *International Journal of Science Education*, *32*(18), 2413-2428. * https://doi.org/10.1080/09500690903431561

Zhang, D. y Campbell, T. (2011). The psychometric evaluation of a three-dimension elementary science attitude survey. *Journal of Science Teacher Education*, *22*, 595-612. * https://doi.org/10.1007/s10972-010-9202-3

Systematic review of attitude toward science instruments (2004-2016)

Radu Bogdan Toma
Departamento de Didácticas Específicas. Facultad de Educación. Universidad de Burgos. Burgos, España.
rbtoma@ubu.es

The purpose of this study was to review the psychometric properties of attitude toward science tools that have been published in the last thirteen years in order to determine whether educational proposals that are being promoted internationally for the improvement of science education are complemented with the development of valid and reliable self-report tools. This analysis, which draws on and update Blalock et al.'s (2008) work, may help researchers to choose adequate data collection instruments according to their study needs.

To this end, a systematic review following the procedure described by Bennett, Lubben, Hogarth and Campbell (2005) and the PRISMA statement (Liberati et al., 2009) was performed. Potentially relevant articles were retrieved from two databases, mainly *Web of Science Core Collection* and *Science Direct*, and using a snowball technique, which consists in examining the reference list of selected articles for relevant studies not identified through the databases. Searches were restricted by research category (Education and educational research), document type (articles), language (English and Spanish) and year of publication (2004-2016). Articles were deemed relevant according to the following inclusion criteria: (I) instrument development and validation studies; (II) focused on the attitudes towards science construct; (III) developed through classical test theory; and (IV) quantitative in nature.

The psychometric properties of each instrument was assessed using a scoring rubric consisting of four parts and with a score ranging from 0 to 27: (I) Theoretical framework –to what extent the tool was developed according to existing attitudinal theories–; (II) Reliability –whether internal consistency, test-retest and standard measurement error results are reported–; (III) Validity –whether content, construct, convergent, concurrent, discriminant, discriminative and predictive validity results are reported–; and (IV) Dimensionality –whether data was analyzed according to the unidimensionality or multidimensionality of the tool–.

A total of 14 studies providing information on the development and validation of 15 instruments were retained for in-depth analysis. Overall, 9 (60 %) instruments scored less than half of the total rubric score (<13). A total of 11 (73 %) instruments were developed without a theoretical framework. In terms of reliability, no study reported psychometric evidence for test-retest or information about the standard error of measurement, and only four (27 %) of the tools reported high levels of reliability. As for the validity, up to 9 (60 %) instruments reported only one or two psychometric evidences (mostly content or construct validity), and no tool was subjected to more than four psychometric validity tests.

These results indicate that Blalock et al.'s (2008) and Munby's (1983) criticisms are still valid for current attitudes towards science instruments. These results call into question the confidence that can be placed in the results derived from studies using instruments whose validity and reliability are at stake, and calls for a reconceptualization of the attitude toward science construct and the adoption of rigorous validation procedures to develop valid and reliable measurement tools that would help support (or refute) the assumptions and consensus reached in this line of research.