ENSEÑANZA DE LAS CIENCIAS, 37-2 (2019), 127-149 https://doi.org/10.5565/rev/ensciencias.2632
Investigaciones didácticas ISSN (impreso): 0212-4521 / ISSN (digital): 2174-6486

Análisis de las pruebas externas de evaluación de la competencia científico-tecnológica de 6.º de Educación Primaria en España (2016)

Analysis of the external assessment of the scientific-technological competence in 6th grade of Primary Education (2016)

Tobías Martín-Páez
Departamento de Didáctica de las Ciencias Experimentales, Universidad de Granada, Granada, España
tmartin@ugr.es

Javier Carrillo-Rosúa
Departamento de Didáctica de las Ciencias Experimentales, Universidad de Granada, Granada, España
Instituto Andaluz de Ciencias de la Tierra (CSIC-UGR),Armilla, España
fjcarril@ugr.es

José Luís Lupiáñez-Gómez
Departamento de Didáctica de las Matemáticas, Universidad de Granada, Granada, España
lupi@ugr.es

José Miguel Vílchez-González
Departamento de Didáctica de las Ciencias Experimentales, Universidad de Granada, Granad, España
jmvilchez@ugr.es

RESUMEN • Este artículo presenta un análisis documental en el que se analizan las pruebas externas de evaluación de la competencia científica y tecnológica elaboradas para el sexto curso de Educación Primaria desde tres puntos de vista. Primero, su ajuste y adecuación a las variables de los marcos teóricos de referencia. Segundo, caracterizar la presencia de la competencia científica y tecnológica en las pruebas y, tercero, determinar las relaciones entre las variables y la competencia evaluada. Se han analizado 169 actividades mediante triangulación entre expertos. Los resultados muestran debilidad en el ajuste al marco teórico, y que la competencia aplicada exige escasa interpretación de datos y diseño de investigaciones. Finalmente, existe una relación simétrica y positiva entre demandas cognitivas y nivel competencial.

PALABRAS CLAVE: Evaluación educativa; Alfabetización científica; Educación basada en competencias; Educación Primaria; Pruebas externas.

ABSTRACT • This article presents a documentary analysis focused on the external tests assessing the scientific and technological competence in the sixth year of Primary Education from three different points of view. Firstly, we pay attention to its adjustment and adaptation to the variables of the reference theoretical frameworks; secondly, we characterize the presence of scientific and technological competence in the tests and, thirdly, we determine the relationships between the variables and the evaluated competence. 169 activities have been analyzed through triangulation among experts. The results show that the adjustment to the theoretical framework is weak and that the applied competence requires scarce interpretation of data and research design. Finally, there is a symmetrical and positive relationship between cognitive demands and competence level.

KEYWORDS: Educational assessment; Scientific literacy; Competence-based education; Elementary school; External assessment.

Recepción: marzo 2018 • Aceptación: noviembre 2018 • Publicación: junio 2019

Martín-Páez, T., Carrillo-Rosúa, J., Lupiáñez-Gómez, J. L., y Vílchez-González, J. M. (2019). Análisis de las pruebas externas de evaluación de la competencia científico-tecnológica de 6.º de Educación Primaria en España (2016). *Enseñanza de las ciencias, 37*(2), 127-149. https://doi.org/10.5565/rev/ensciencias.2632

INTRODUCCIÓN

En la sociedad actual se interactúa continuamente con información de carácter científico y tecnológico y, como señala Sjøberg (2015), los avances actuales en campos como la medicina, los nuevos medios de comunicación y los artefactos que conviven a nuestro alrededor ponen de manifiesto que ciencia y tecnología avanzan a un ritmo trepidante y sorprendente. En este contexto cobra especial importancia la alfabetización científica y tecnológica en los escolares y, de hecho, son ya frecuentes las reflexiones curriculares acerca de lo prioritario de una formación de calidad en estos ámbitos (Hackling, Ramseger y Chen, 2017). El término *alfabetización científica* es complejo y admite distintas definiciones (Bybee, 1997). Por ello, Roberts (2007) propone dos visiones diferentes, una que enfatiza los productos y procesos de la ciencia y otra que relaciona la ciencia con contextos relevantes en la sociedad y asuntos de la vida cotidiana. Feinstein (2011) pone de manifiesto el continuo debate sobre la interpretación de estas visiones, así como la necesidad de incluir aspectos adicionales de carácter social. Actualmente se han visto superadas por nuevas propuestas que implican no solo la relación entre ciencia, tecnología y sociedad, sino el desarrollo de ciudadanos críticos y activos con un desempeño científico en la sociedad (Sjöström, Eilks y Zuin, 2016). Se asume la definición establecida por la OCDE (2017) según la cual la alfabetización científica es «la capacidad de involucrarse en temas relacionados con la ciencia y con las ideas de la ciencia, como un ciudadano reflexivo. Una persona con conocimientos científicos está dispuesta a participar en un discurso razonado sobre ciencia y tecnología» (p. 24). Esta visión de la alfabetización científica se contrapone a otra más clásica y academicista, compartida por una parte de la sociedad, centrada principalmente en la adquisición de conocimientos científicos con el fin de familiarizar a los estudiantes con las teorías, conceptos y procesos científicos necesarios para continuar su avance a través de las sucesivas etapas educativas, y que se ajusta mejor a las nuevas definiciones del término.

Las autoridades educativas, a través de sus legislaciones, tratan de dar respuesta a esta necesidad de alfabetizar científicamente a sus ciudadanos. Así, en muchos países, durante los últimos treinta años, la enseñanza de las ciencias ha sido orientada hacia la alfabetización científica de los futuros ciudadanos (Furió, Vilches, Guisasola y Romo, 2001), implementando diversas respuestas que originan distintas estrategias sobre cómo participar en su desarrollo y sobre el tipo de currículum que lo propicia (Fensham, 1985; Hodson y Reid, 1988). En la tabla 1 se recogen algunos ejemplos de estas iniciativas institucionales. La importancia de estas propuestas curriculares también se sostiene en la convicción de que esa formación científico-tecnológica puede minimizar el impacto de algunas brechas sociales en la ciudadanía (Berube, 2014; Babaci-Wilhite, 2016).

Tabla 1.
Iniciativas institucionales que enfatizan la relevancia de la alfabetización científica

Nombre	*Autor*	*Objetivo*
El proyecto DeSeCo (Definition and Selection of Competencies)	Organización para la Cooperación y el Desarrollo Económicos (OCDE)	Proporcionar un marco conceptual sólido que establezca los objetivos que debe alcanzar cualquier sistema educativo que pretenda fomentar la educación a lo largo de toda la vida.
PISA (Programme for International Student Assessment)	Organización para la Cooperación y el Desarrollo Económicos (OCDE)	Evaluar la formación de los alumnos cuando llegan al final de la etapa de enseñanza obligatoria, hacia los 15 años.
El Movimiento de Educación para Todos	Conferencia Mundial de Educación de 1990 (UNESCO, el PNUD, el FNUAP, UNICEF y el Banco Mundial)	Dar una educación básica de calidad a todos los niños, jóvenes y adultos.
El Proyecto 2061	Asociación Americana para el Avance de la Ciencia (AAAS)	Alcanzar una adecuada instrucción en las Ciencias, Matemáticas y Tecnología.

La competencia científica y tecnológica es la respuesta educativa a la alfabetización científica. En el caso concreto de la evaluación por competencias, las instituciones educativas han tratado de dar respuesta a esta complejidad con diversas propuestas, llegando a ser una temática de creciente interés social con frecuentes referencias en los medios de comunicación (e.g. Álvarez, 2016; Coughlan, 2016; Santos, 2016; Leguilloux, 2018). No solo se han replanteado los elementos de la evaluación (competencias, criterios de evaluación, estándares de aprendizaje), sino que también se han redefinido otros aspectos, como quién evalúa (interna o externamente) o cuál debe ser el objetivo de esta fase del proceso de enseñanza-aprendizaje (diagnóstico, regulación o promoción) (Sanmartí, 2007; Black y William, 2018; Stanley, MacCann, Gardner, Reynolds y Wild, 2009).

Por su parte, la evaluación externa suele tener carácter censal y se implementa mediante pruebas de papel y lápiz administradas colectivamente, pues a las dificultades anteriores se añade la necesidad de evaluar a una gran población de estudiantes. Además, en el caso de la evaluación de la alfabetización científica, en estas pruebas estandarizadas suelen presentarse escasos conocimientos sobre la naturaleza de la ciencia (Lau, 2009). Algunos ejemplos de estas evaluaciones externas internacionales son el *Programme for International Student Assessment* (PISA) (OCDE, 2018) o el estudio *Trends in International Mathematics and Science Study* (TIMSS) (Martin, Mullis, Foy y Hooper, 2015).

EL CASO DE ESPAÑA

El sistema educativo español aborda la evaluación por competencias en su desarrollo legislativo, tanto como parte de los elementos curriculares (recoge siete competencias clave, entre ellas la competencia matemática y competencias básicas en ciencia y tecnología) como por la implantación de programas de evaluación externa. El Real Decreto 126/2014, por el que se establece el currículo básico de Educación Primaria (MECD, 2014), recomienda la elaboración de portafolios y la observación directa durante el proceso de enseñanza-aprendizaje, de forma combinada con sistemas de rúbrica para la evaluación del alumnado. También establece la evaluación externa como elemento que permite obtener información del sistema educativo y distribuir de manera más eficiente los recursos disponibles entre las comunidades educativas; para ello se han establecido dos momentos, uno en 3.º y otro en 6.º de Educación Primaria, aunque la competencia científico-tecnológica se evalúa solo en 6.º curso.

ESTADO DE LA CUESTIÓN

Los programas de evaluación externa generan una gran cantidad de datos sobre el desempeño escolar del alumnado, sus características socioeconómicas, etc. El tratamiento de esta cantidad de datos es complicado y por este motivo se pueden encontrar trabajos, como el de Jerrim, López-Agudo, Marcenaro-Gutiérrez y Shure (2017), que tratan de generar una mejor comprensión de las bases de datos internacionales de educación a gran escala y promover mejores prácticas en su uso.

El tratamiento de estos datos permite la generación de investigaciones, como las realizadas por Vidal, Díaz y Jarquín (2004), que establecen relaciones entre el desempeño escolar y determinados factores demográficos. Este tipo de investigaciones son de reconocido interés debido a que brindan información sobre los factores que influyen en el rendimiento escolar del alumnado y permiten a las instituciones educativas gestionar mejor sus recursos al orientarlos a determinados colectivos (Benavides, León y Etesse, 2014) o fomentar innovaciones y mejoras en determinadas etapas educativas (Pholphirul, 2017).

Pero hay que tener en cuenta que estas pruebas presentan limitaciones y por eso han sido analizadas en publicaciones centradas en discutir los defectos en la elaboración de las pruebas o sus mar-

cos teóricos. Solano-Flores, Contreras-Niño y Backhoff-Escudero (2006), por ejemplo, estudian las limitaciones lingüísticas de las pruebas PISA al ser elaboradas en inglés y tener que traducirse a las diversas lenguas de los países destinatarios. Las debilidades de sus marcos teóricos han sido analizadas en publicaciones como la de Acevedo (2005) en Ciencias, y Caraballo, Rico y Lupiáñez (2013) en Matemáticas.

En este trabajo se realiza un análisis descriptivo expostfacto que se centra en la descripción y el análisis de las pruebas de evaluación externa ya implementadas. Se pueden encontrar investigaciones similares previas, como la de yus *et al.* (2013), en la que se analizan las pruebas PISA centrando la atención en la demanda de la competencia científica y poniendo de manifiesto que se utilizan en mayor medida las capacidades que implican un menor nivel cognitivo.

OBJETIVOS

Los objetivos de esta investigación se orientan hacia el análisis de las pruebas de evaluación externa de la competencia científico-tecnológica elaboradas para 6.º curso de Educación Primaria del año 2016:

1. Comprobar si las pruebas de evaluación externas se ajustan a sus marcos de referencia.
2. Caracterizar la presencia de la competencia científica y tecnológica en estas pruebas.
3. Determinar si en esas pruebas existen relaciones entre las variables definidas en los marcos de referencia para evaluar la competencia científica y tecnológica.

METODOLOGÍA DE INVESTIGACIÓN

En esta investigación se desarrolla un análisis documental de unas pruebas de evaluación externas. Debido a las debilidades de las metodologías cuantitativas en este tipo de investigaciones sociales (Mayring, 2017), los datos han sido recogidos, filtrados y consensuados a través de una triangulación (Mihladiz y Dogan, 2017) realizada por tres de los autores, expertos en Didáctica de las Ciencias Experimentales, a través de un sistema de variables que parten de las definiciones dadas desde el marco general (MG) de la evaluación final de Educación Primaria (MECD, 2015a) y la OCDE (OCDE, 2017). El análisis ha generado unidades de información cuantificadas (Martin, Gómez y Verde, 2017), permitiéndonos su análisis estadístico mediante frecuencias relativas, el estadístico V de Cramer y los estadísticos Lambda y Tau de Goodman y Kruskal (1963).

UNIDADES DE INFORMACIÓN

Se han analizado las cinco pruebas de evaluación externa de la competencia científica y tecnológica para 6.º curso de Educación Primaria disponibles del curso 2015-16 (tabla 2).

Tabla 2.
Pruebas analizadas y su codificación

Codificación	*Autor*	*Ámbito de aplicación*
p-me	Ministerio de Educación, Cultura y Deporte	Territorio MECD, Ceuta, Melilla, Madrid, La Rioja y Castilla y León
p-cm	Consejería de Educación, Cultura y Deportes (Castilla-La Mancha)	Castilla la Mancha

Codificación	*Autor*	*Ámbito de aplicación*
p-ga	Consellería de Cultura, educación e ordenación universitaria (Xunta de Galicia)	Galicia
p-mu	Consejería de Educación y Universidades (Murcia)	Murcia
p-na	Departamento de Educación (Gobierno de Navarra)	Navarra

Nota: en adelante nos referiremos a las pruebas por su codificación.

No se han analizado más pruebas de comunidades autónomas bien porque no han liberalizado las pruebas (e. g. Islas Canarias) o bien porque no las han implementado (e. g. Andalucía o Cataluña).

Cada una de estas pruebas consta de unidades de evaluación, y estas, a su vez, de un estímulo (información sobre una situación problemática contextualizada) con preguntas asociadas (tabla 3).

Tabla 3.
Cuadro resumen de las pruebas analizadas

Pruebas	*Estímulos (n.º de preguntas de cada estímulo)*	*Preguntas*
p-me	Brote de sarampión (3) Cómo montar un acuario (6) Un hueso roto (4) Ruta en bicicleta (6) Trucos de espías (6) ¿Por qué desaparecieron los dinosaurios? (5) Obesidad infantil (6) Construye online (4)	40
p-cm	Brote de sarampión (3) Cómo montar un acuario (4) Un hueso roto (4) Ruta en bicicleta (3) Trucos de espías (3) ¿Por qué desaparecieron los dinosaurios? (3) Obesidad infantil (3) Construye en línea «online» (2)	25
p-ga	Viaxe de fin de curso (6) Rodeados de Natureza (6) Somos o que comemos, pero tamén como nos mvemos! (6) Hai un apagamento… E agora, que? (5) Parada na cidade (6) De volta na casa (7)	36
p-mu	Una llegada accidentada (11) La vida en el planeta Tierra (8) Nuestros inventos (11) Salud (7)	37
p-na	Proyecto intercultural (6) Reciclaje (6) En el albergue (6) Bodas de oro (7) Excursión a San Sebastián (6)	31

De esta manera, la muestra la constituyen las actividades de las pruebas de evaluación final de 6.º de Educación Primaria que han sido diseñadas, implementadas y liberalizadas para el curso académico 2015-16 en España. Consta, pues, de 169 actividades de evaluación.

Descripción de las variables

Para definir las variables de análisis se utilizan dos marcos de referencia: el MG, del que se extraen las dimensiones de la competencia científico-tecnológica (contextos, contenidos y demandas cognitivas) y el formato de pregunta, y el establecido por la OCDE, según el cual el desarrollo de esta competencia se relaciona con el desarrollo de tres subcompetencias: «Explicar fenómenos científicamente», «Evaluar y diseñar investigaciones científicas» e «Interpretar datos científicamente» (OCDE, 2017). En este estudio no ha sido considerada la variable actitud, presente en el MG, pues en la propia norma se recomienda su evaluación mediante la observación directa del alumnado.

La tabla 4 muestra las variables utilizadas para dar respuesta al primero de nuestros objetivos y los posibles valores de estas; entre paréntesis se indica, si procede, la frecuencia relativa recomendada para estos valores desde el MG.

Tabla 4.
Dimensiones de la competencia
científico-tecnológica y formato de pregunta

Variables	*Valores*
Contexto	Personal
	Escolar
	Social
	Artístico-humanístico
	Otras contextualizaciones
Contenido	Los seres vivos (20 %)
	El ser humano y la salud (20 %)
	La materia y la energía (30 %)
	La tecnología, los objetos y las máquinas (30 %)
	Acceso e identificación (20 %)
	Comprensión (20 %)
Demanda cognitiva	Aplicación (20 %)
	Análisis (20 %)
	Síntesis y creación (15 %)
	Juicio y valoración (5 %)
	No observada (0 %)
Formato de pregunta	Cerrada (mínimo 40 %)
	Semiconstruida
	Construida (mínimo 20 % entre todas)
	Abierta

Contexto

Los valores de la variable contexto se definen como:

a) Personal: problemas o desafíos a los que podría enfrentarse el alumno relacionados con él mismo, su familia o su grupo de amigos.
b) Escolar: situaciones relacionadas con la vida escolar y el grupo o grupos de compañeros.
c) Social: situaciones referidas al barrio, a la localidad o a la sociedad en general.
d) Artístico y humanístico: relacionado con la aplicación de la ciencia y la tecnología al mundo artístico y de las Ciencias Sociales. El término *ciencias sociales* es muy amplio, por lo que se entenderá como la aplicación de la ciencia en las disciplinas humanísticas.
e) Otras contextualizaciones: actividades que no están contextualizadas y tampoco presentan relación con el contexto de la situación problemática o una actividad anterior. También se incluyen en este apartado aquellas actividades que no presentan relación con ninguno de los contextos establecidos en el MG.

Los valores «Personal», «Escolar» y «Social» coinciden con la descripción establecida desde el MG. El valor «Artístico y humanístico» ha sido ampliado para que incluya la aplicación de la ciencia en las disciplinas humanísticas. El valor «Otras contextualizaciones» ha sido creado *ad hoc* para agrupar aquellas actividades que no se pueden asociar a ninguno de los contextos definidos en el MG.

Contenido

Para analizar esta variable se utilizaron los bloques de contenido descritos por el MECD en el RD 126/2014 (MECD, 2014), a excepción del bloque «Iniciación a la actividad científica». La exclusión se debe a que el MG establece que este bloque se encuentra incluido en el resto de bloques de contenido en 6.º de Educación Primaria. De este modo queda definida la variable contenido y sus valores: «Los seres vivos», «El ser humano y la salud», «La materia y la energía» y «La Tecnología, los objetos y las máquinas». Los valores de esta variable no se describen explícitamente en este artículo ya que se encuentran definidos en el RD 126/2014 (MECD, 2014).

Demanda cognitiva

La variable demanda cognitiva ha sido analizada a partir de la caracterización realizada en el MG (figura 1):

Fig. 1. Procesos cognitivos (elaboración propia).

a) Acceso e identificación: acciones de recordar y reconocer hechos, conceptos y relaciones; características o propiedades de organismos, materiales o dispositivos; usos de equipos y procedimientos; usar vocabulario científico-tecnológico, abreviaturas, unidades, símbolos y escalas. Buscar y seleccionar información relevante sobre los contenidos.
b) Comprensión: describir o identificar descripciones de propiedades, estructuras, funciones de organismos, materiales o dispositivos, y las relaciones entre estos y los procesos o fenómenos. Dar ejemplos de organismos, materiales, dispositivos o procesos que tienen determinadas características. Explicar hechos y conceptos con los ejemplos adecuados.
c) Aplicación: identificar o describir semejanzas o diferencias entre grupos de organismos, materiales, dispositivos o procesos. Distinguirlos y clasificarlos. Vincular el conocimiento de un concepto subyacente científico-tecnológico a propiedades, comportamiento o uso, observado o inferido, de estos.
d) Análisis: utilizar diagramas u otros modelos para demostrar el conocimiento de conceptos científico-tecnológicos, ilustrar un proceso o sistema o para encontrar soluciones a problemas. Utilizar el conocimiento de conceptos científico-tecnológicos para interpretar información relevante. Explicar una observación utilizando un concepto o principio científico-tecnológico, con el vocabulario científico-técnico adecuado. Presentar información de forma coherente, ordenada y clara.
e) Síntesis y creación: obtener y analizar datos y otras informaciones, extraer conclusiones, extrapolar lo comprendido a nuevas situaciones, en contextos poco habituales, desarrollar hipótesis. Utilizar métodos propios de observación. Diseñar y realizar experiencias sencillas y pequeñas investigaciones.
f) Juicio y valoración: emitir opiniones argumentadas aplicando conocimientos científico-tecnológicos. Valorar aspectos relacionados con la ciencia y la tecnología y sus aplicaciones (p. 95).

Formato de pregunta

El formato de pregunta ha sido caracterizado mediante la clasificación establecida en el MG:

a) Preguntas de respuesta cerrada, bajo el formato de elección múltiple, en las que solo una opción es correcta y las restantes se consideran erróneas.
b) Preguntas de respuesta semiconstruida, que incluyen varias preguntas de respuesta cerrada dicotómicas o solicitan al alumnado que complete frases o que relacione por medio de flechas diferentes términos o elementos.
c) Preguntas de respuesta construida, que exigen el desarrollo de procedimientos y la obtención de resultados.
d) Preguntas de respuesta abierta, que admiten respuestas diversas, las cuales, aun siendo correctas, pueden diferir unas de otras.

Esta variable está relacionada con las características generales de la prueba y su presencia implicará respuestas con mayor objetividad en la corrección (respuestas cerradas), o aquellas que darán mayor información al evaluador (argumentación).

Subcompetencia

Para el segundo objetivo se utilizan las subcompetencias asociadas a la competencia científico-tecnológica por la OCDE (2017) y asumidas desde el MG. La figura 2 muestra los valores de esta variable:

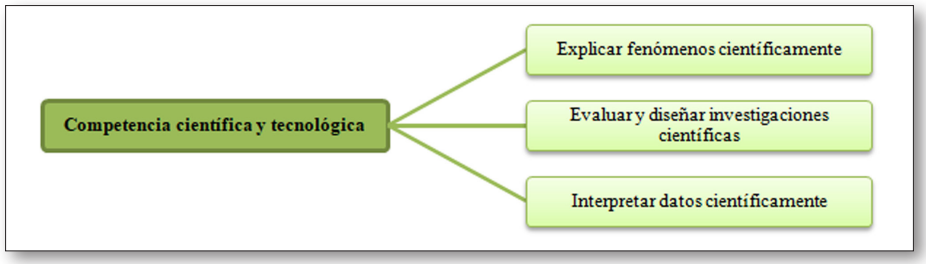


Fig. 2. Competencia científica y tecnológica (elaboración propia).

Las definiciones de estos valores serían:

a) Explicar fenómenos científicamente: tiene que ver con recordar y utilizar teorías, ideas, información y hechos; ofrecer explicaciones científicas, lo que requiere una comprensión de cómo este conocimiento se ha creado y el nivel de confianza que se puede tener en las afirmaciones científicas; conocer las formas y procedimientos estándares que se utilizan en la investigación científica para obtener dichos conocimientos y comprender su papel en la justificación de los conocimientos científicos.

b) Evaluar y diseñar investigaciones científicas: nos permite conocer los procedimientos básicos usados en el ámbito científico y la función de estos procedimientos a la hora de evaluar los nuevos avances científicos, permitiendo definir cómo una problemática de carácter científico puede ser resuelta.

c) Interpretar datos científicamente: nos permite dar argumentaciones basadas en datos científicos e interpretar si los datos dados en los argumentos de otras personas, y por lo tanto sus argumentos, son incorrectos.

El desarrollo de estas subcompetencias supone el de la competencia científica y tecnológica (OCDE, 2017). Esta afirmación, unida al hecho de que otros trabajos ponen de manifiesto su validez en este tipo de estudios (Yus *et al.*, 2013), es lo que permite considerar que son adecuadas para este estudio.

INSTRUMENTO DE RECOGIDA DE DATOS Y PROCEDIMIENTO

Como instrumento para la recogida de datos se ha utilizado una parrilla de doble entrada. En la primera fila se situaron las variables analizadas, y en la segunda, los valores asociados a cada una de ellas. A partir de la tercera fila, en la primera columna se encuentran las referencias de las actividades, y el resto de columnas se dejaron en blanco para anotar los resultados de la observación. En la tabla 5 se muestra un fragmento de esta:

Tabla 5.
Fragmento de la parrilla de doble entrada

	Contexto				Contenido		…
	Personal	Escolar	Social	Artístico-humanístico	Los seres vivos	El ser…	…
P-ME1*							
P-ME2							

* P-ME1: Prueba del Ministerio de Educación, Cultura y Deporte, actividad 1.

En cuanto al procedimiento, se ha optado por un grupo compuesto por tres expertos, investigadores del Departamento de Didáctica de las Ciencias Experimentales de la Universidad de Granada.

Cada experto valoró, de manera independiente, las actividades de cada una de las pruebas en función de las variables del estudio. Una vez obtenidos los primeros resultados, se procedió a su depuración a través de un proceso previamente establecido, que consistía en:

a) Una primera reunión para intentar establecer un consenso.
b) De no alcanzarlo, se realiza otra valoración individual.
c) En otra nueva reunión se intenta concretar una valoración conjunta.
d) En caso de no alcanzar aún un acuerdo, se acude a un juez experto externo.

El procedimiento finalizó en la segunda reunión, apartado *c*), aunque fue necesario definir una serie de criterios para llegar a consensuar las observaciones. La figura 3 muestra estos criterios agrupados por variables:

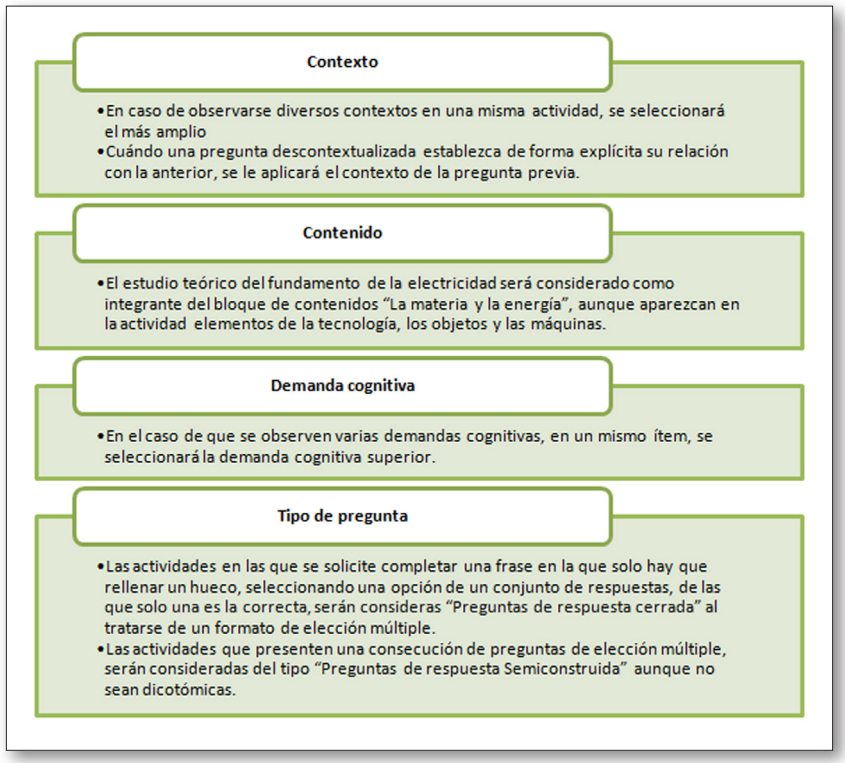


Fig. 3. Criterios de valoración (elaboración propia).

Para analizar los datos se utilizó el software SPSS v22, considerando frecuencias relativas y tablas de variables cruzadas para favorecer la descripción de las pruebas, el estadístico V de Cramer para establecer la relación estadística entre las variables y los estadísticos Lambda y Tau de Goodman y Kruskal para reflejar la reducción del error al pronosticar una variable a partir de otra.

RESULTADOS Y DISCUSIÓN

Los resultados y la discusión se presentarán atendiendo a los objetivos de la investigación.

COMPROBAR SI LAS PRUEBAS DE EVALUACIÓN EXTERNAS SE AJUSTAN A SUS MARCOS DE REFERENCIA

A continuación, se muestran las frecuencias relativas asociadas a cada una de las variables analizadas y las frecuencias prescritas por sus marcos de referencia, que son el MG, la prueba modelo (MECD, 2015c) o ambos, dependiendo del caso.

Contexto

La variable contexto presenta las frecuencias relativas reflejadas en la figura 4.

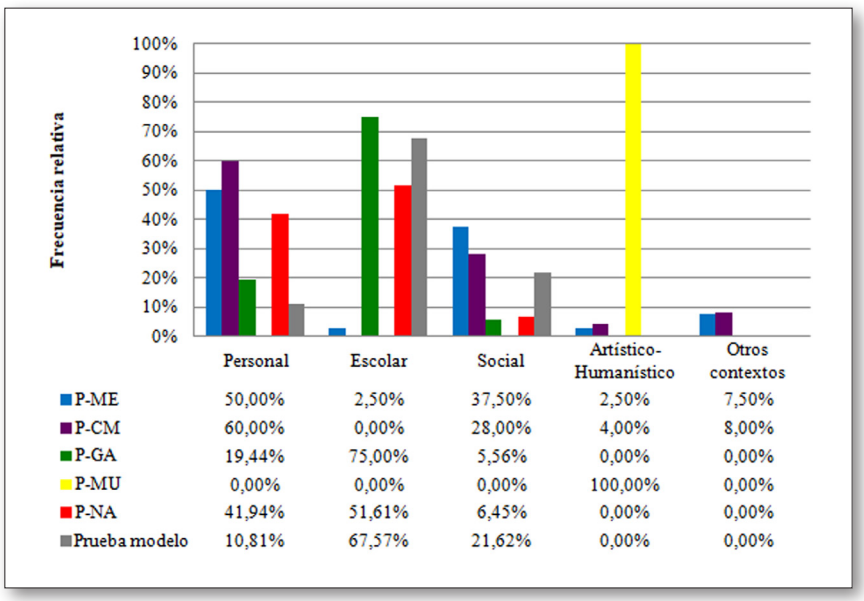


Fig. 4. Frecuencias relativas de la variable Contexto en las pruebas analizadas y en la prueba modelo.

El análisis de esta variable pone de manifiesto que hay diferencias notables entre las distintas pruebas analizadas y entre estas y la prueba modelo:

a) El valor «Personal» tiene una presencia mayor en las pruebas P-ME, P-CM, P-GA y P-NA y no se presenta en la prueba P-MU.
b) El valor «Escolar» tiene una presencia importante en las pruebas P-GA y P-NA, siendo menor su presencia en la prueba P-ME y no apareciendo en las pruebas P-CM y P-MU.
c) El valor «Social» se da con una mayor frecuencia en las pruebas P-ME y P-CM, en menor medida en las pruebas P-GA y P-NA y no se utiliza en la prueba P-MU.
d) El valor «Artístico-humanístico» se da mayoritariamente en la prueba P-MU ya que todas sus actividades se basan en este contexto; también encontramos una leve presencia en las pruebas P-ME y P-CM, mientras que no se encuentra en las pruebas P-GA y P-NA.
e) El valor «Otras contextualizaciones» solo aparece en las pruebas P-ME y P-CM.

La distribución de los valores de la variable no está establecida en el MG, aunque existe una normativa (OCDE, 2017; MECD, 2014 y 2015b) que fomenta la diversidad de contextos cercanos al alumnado para el desarrollo y evaluación de las competencias. Bajo este supuesto, los contextos «Personal» y «Escolar»

deberían tener más presencia que el resto. Al utilizar diferentes contextualizaciones se rompe la brecha entre el contexto real en el que se produce la evaluación y otros contextos (situaciones de la vida cotidiana), de forma que el alumnado podrá demostrar mejor su desarrollo competencial en otras situaciones (Kuglitsch, 2015). Además, se pone de manifiesto la descontextualización de algunos de los ítems y la lejanía de algunos de los contextos respecto a situaciones reales, como es el caso de la prueba P-MU, en la que se pierde parte de su significación para fomentar una posible motivación incluyendo un contexto fantástico.

Contenido

En la figura 5 se muestran las frecuencias relativas de esta variable en cada una de las pruebas analizadas.

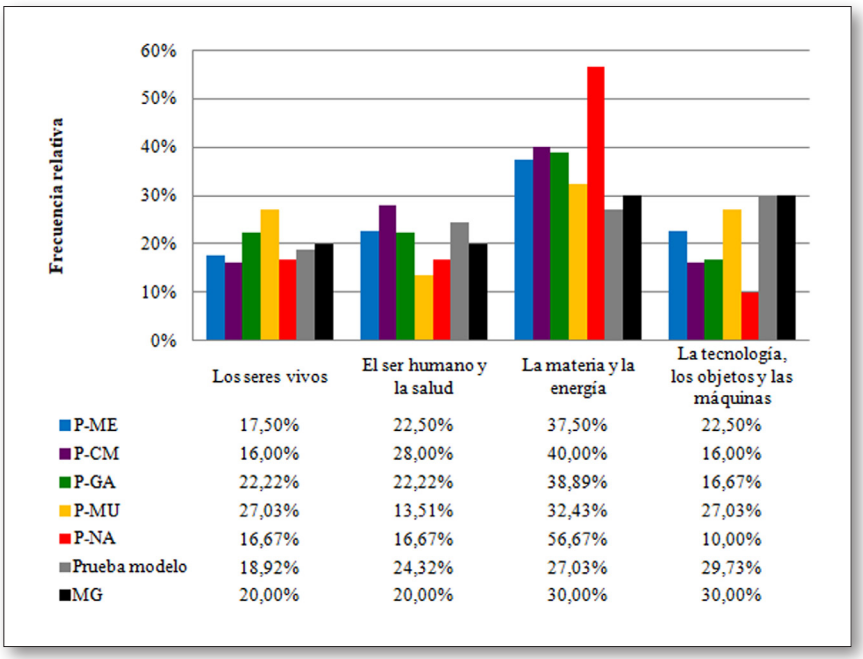


Fig. 5. Frecuencias relativas de la variable Contenido en las pruebas analizadas, prueba modelo y MG.

La variable Contenido presenta las siguientes frecuencias relativas:

a) El valor «Los seres vivos» presenta frecuencias superiores en las pruebas P-MU y P-GA (27,03 y 22,22 %) a las de la prueba modelo (18,92 %) y a la prescrita desde el MG (20,00 %) y una presencia menor en el resto de pruebas analizadas.

b) El bloque de contenidos «El ser humano y la salud» tiene una frecuencia relativa menor a la de la prueba modelo y el MG en las pruebas P-MU (13,51`%) y P-NA (16,67 %), intermedia entre estos dos referentes en la prueba P-GA y superior en la prueba P-CM.

c) El valor «La materia y la energía» tiene frecuencias superiores al 30 % e inferiores al 40 % en las pruebas P-ME, P-CM, P-GA y P-NA y una frecuencia del 56,67 % en la prueba P-NA. Todas las pruebas presentan una frecuencia mayor del valor establecido en la prueba modelo, en la que el 27,03 % de las actividades presentan este valor, y al 30,00 % prescrito desde el MG.

d) En el bloque de contenidos «La tecnología, los objetos y las máquinas», en el que la prueba modelo y el MG establecen el 29,73 y el 30,00 % respectivamente, todas las pruebas presentan frecuencias relativas inferiores, aunque la P-MU es la que más se acerca, con el 27,03 %, y la P-NA la que más se aleja, con el 10,00 %.

e) La prueba modelo se adecúa en mayor medida a la distribución de los bloques de contenidos establecida por el MG que las pruebas analizadas.

El MG demanda una mayor presencia de los bloques de contenido «La materia y la energía» y «La tecnología, los objetos y las máquinas» que la observada en la mayoría de las pruebas. Trabajos como el de Kwak (2017) ponen de manifiesto las dificultades del alumnado con estos contenidos y la necesidad de la promoción y la innovación en estos en los diseños curriculares.

En España, la normativa estatal permite la adecuación de contenidos a las singularidades de las comunidades autónomas, incluso a nivel de centro o aula, a través de las diversas concreciones curriculares, lo que se podría trasladar a esas pruebas. Esto, que en principio podría parecer aconsejable, también puede arrastrar inconsistencias, pues el alumnado podría evidenciar diferente rendimiento en pruebas locales que en globales (Anagnostopoulou, Hatzinikita, Christidou y Dimopoulos, 2013; Hatzinikita, Dimopoulos y Christidou, 2008).

Formato de pregunta

Las frecuencias relativas de la variable Formato de pregunta pueden ser observadas en la figura 6.

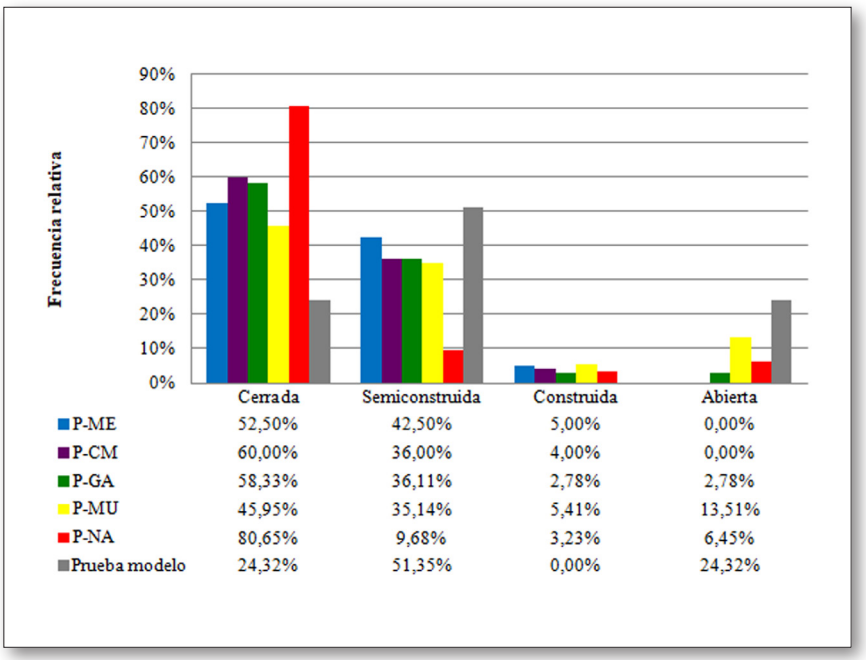


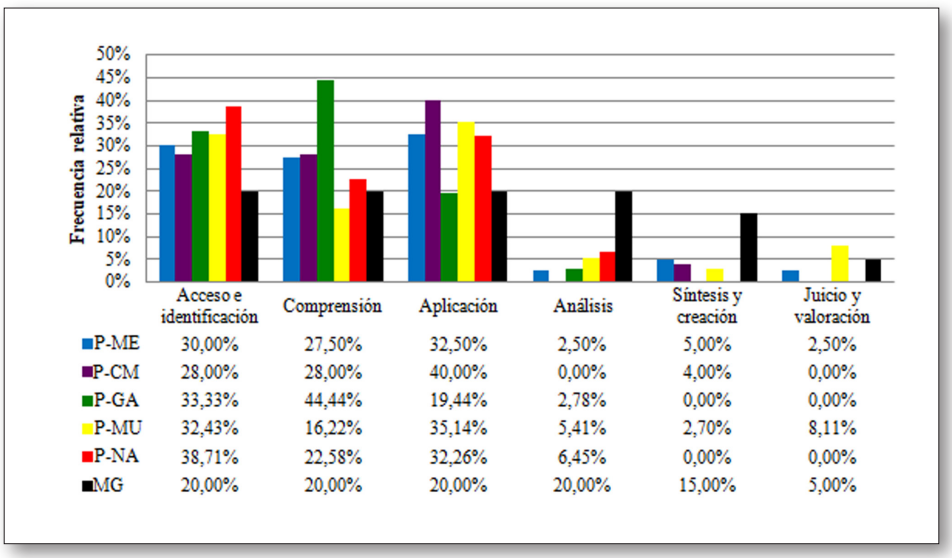
Fig. 6. Frecuencias relativas de la variable Tipo de pregunta de las pruebas analizadas y de la prueba modelo.

Según lo preestablecido por el MG, las preguntas cerradas deben ser como mínimo el 40,00 %, y el resto de tipologías de interrogantes, en conjunto, deben ser un mínimo del 20,00 %. Las pruebas analizadas se ajustan a esta demanda, lo que no ocurre en la prueba modelo.

El formato de pregunta limita lo que se puede llegar a evaluar en una prueba escrita. Por ejemplo, con una pregunta cerrada (entendida como de opción múltiple, como se hace en el MG) nunca se podrá evaluar la capacidad de argumentación del alumnado, limitando así el desarrollo, o la evaluación, de ciertas subcompetencias. En este sentido, una gran diversidad de publicaciones ponen de manifiesto el papel de la argumentación en el desarrollo de la competencia científica y tecnológica (e. g. Jiménez-Aleixandre, 2011; Teixeira y Greca, 2015). Además, existen nuevas tendencias educativas, como la metodología *Argument-driven inquiry* (ADI[1]) (Sampson, Grooms y Walker, 2011), o propuestas de evaluación e intervención (Custodio, Márquez y Sanmartí, 2015; Ouariachi, Olvera-Lobo y Gutiérrez-Pérez, 2017) que se centran especialmente en el desarrollo de dicha competencia. Estas nuevas tendencias educativas permiten el desarrollo y la mejora en la redacción de las estructuras y los contenidos de los argumentos, lo que mejora la comunicación de los conocimientos adquiridos (Pinar y Gülüzar, 2017). Las pruebas analizadas utilizan mayoritariamente las preguntas que no demandan el uso de argumentos y por lo tanto dificultan la evaluación de la competencia científica y tecnológica, dificultad que se ve acrecentada por la diversidad de contextos en los que se hace necesaria la argumentación como solicitan las diversas instituciones educativas (Lee, 2017).

Demanda cognitiva

La figura 7 muestra las frecuencias relativas de la variable Demanda cognitiva en las pruebas analizadas y la demandada desde el MG.



	Acceso e identificación	Comprensión	Aplicación	Análisis	Síntesis y creación	Juicio y valoración
P-ME	30,00%	27,50%	32,50%	2,50%	5,00%	2,50%
P-CM	28,00%	28,00%	40,00%	0,00%	4,00%	0,00%
P-GA	33,33%	44,44%	19,44%	2,78%	0,00%	0,00%
P-MU	32,43%	16,22%	35,14%	5,41%	2,70%	8,11%
P-NA	38,71%	22,58%	32,26%	6,45%	0,00%	0,00%
MG	20,00%	20,00%	20,00%	20,00%	15,00%	5,00%

Fig. 7. Frecuencias relativas de la variable Demanda cognitiva de las pruebas analizadas y del MG.

Estos resultados muestran, respecto a las demandas cognitivas de menor nivel:

a) La demanda cognitiva «Acceso e identificación», demanda cognitiva de nivel inferior, es requerida en valores comprendidos entre el 28,00 y el 38,71 % de las actividades de las pruebas analizadas, presentando frecuencias relativas excesivas, superiores a las demandadas desde el MG.

1. Disponible en línea: <https://www.argumentdriveninquiry.com/>.

b) La demanda cognitiva «Comprensión» presenta también valores superiores a los demandados por el MG en todas las pruebas analizadas, a excepción de la P-MU, que tiene una frecuencia relativa del 16,22 %; además, la P-GA presenta la frecuencia relativa máxima de todas las demandas cognitivas estudiadas.
c) El valor «Aplicación» presenta frecuencias relativas superiores al 30,00 % en las pruebas P-ME, P-CM, P-MU y P-NA y del 19,44 % en la prueba P-GA; el MG establece un frecuencia relativa del 20,00 %, a la que únicamente se aproxima la prueba P-GA.

Respecto a la de más nivel cognitivo, los resultados indican:

a) Las demandas cognitivas «Análisis» y «Síntesis y creación» presentan frecuencias relativas comprendidas entre el 0,00 y el 7,00 % y, en cualquier caso, frecuencias relativas inferiores al 20,00 y al 15,00 %, que son las solicitadas desde el MG.
b) La demanda cognitiva «Juicio y valoración» solo aparece en la prueba P-MU con una frecuencia relativa del 8,11 %, porcentaje ligeramente superior al establecido desde el MG (5,00 %).

El análisis de estas pruebas presenta resultados similares a los obtenidos por Gallardo-Gil *et al.* (2010), Gallardo-Gil, Mayorga-Fernández y Sierra-Nieto (2014) o Yus *et al.* (2013), para las pruebas PISA entre los años 2000 y 2006, en los que también se dan con mayor frecuencia las demandas cognitivas de bajo nivel.

CARACTERIZAR LA PRESENCIA DE LA COMPETENCIA CIENTÍFICA Y TECNOLÓGICA EN ESTAS PRUEBAS

La figura 8 muestra las frecuencias relativas de las subcompetencias implicadas en la competencia científica y tecnológica (OCDE, 2017):

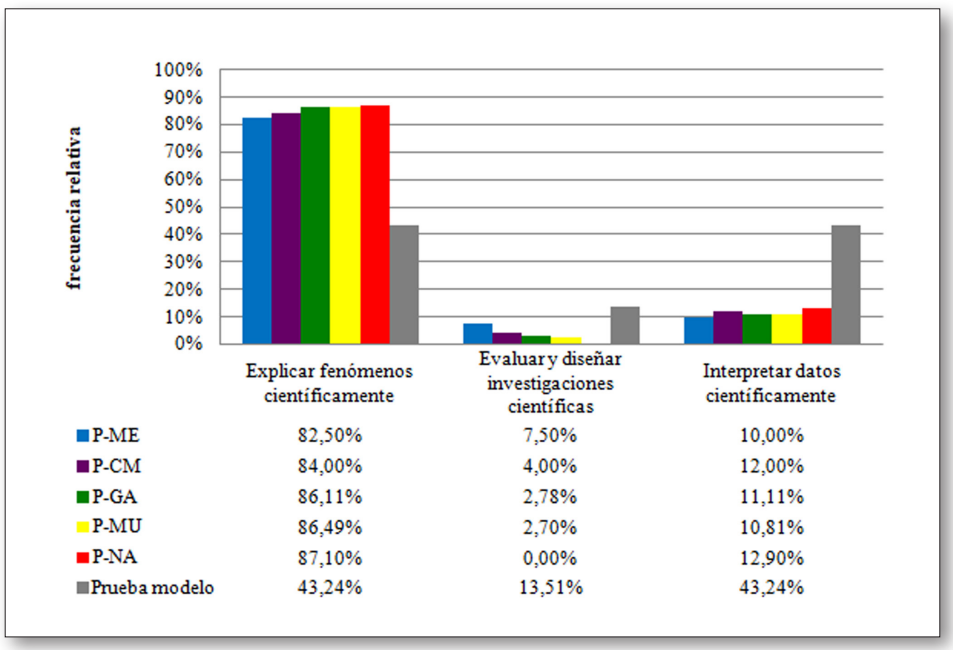


Fig. 8. Frecuencias relativas de la variable Capacidades competenciales de las pruebas analizadas y de la prueba modelo.

Los resultados muestran que:

a) En todas las pruebas analizadas, más del 80,00 % de las actividades de evaluación requerían la subcompetencia «Explicar fenómenos científicamente»; mientras que en la prueba modelo se requería en un 43,24 % de las actividades.
b) La subcompetencia «Evaluar y diseñar investigaciones científicas» es la menos utilizada en todas pruebas, incluso en la prueba modelo, aunque esta prueba es la que presenta una frecuencia relativa mayor.
c) La subcompetencia «Interpretar datos científicamente» presenta frecuencias relativas cercanas al 10,00 % en las pruebas analizadas y del 43,24 % en la prueba modelo.

Se podría sugerir un mayor uso de las subcompetencias «Evaluar y diseñar investigaciones científicas» e «Interpretar datos científicamente», debido a que, según Couso (2014), las prácticas en la educación científico-tecnológica asociadas a la indagación, la modelización y la argumentación aparecen interrelacionadas entre sí. Ejemplos de estas relaciones, y la necesidad de aplicar estas subcompetencias, son la indagación orientada a argumentar (Argument-Driven Inquiry o ADI) o la indagación centrada en modelizar (Model-Based Inquiry o MBI), entre otras (Hernández, Couso y Pintó, 2015).

DETERMINAR SI EN ESAS PRUEBAS EXISTEN RELACIONES ENTRE LAS VARIABLES DEFINIDAS EN LOS MARCOS DE REFERENCIA PARA EVALUAR LA COMPETENCIA CIENTÍFICA Y TECNOLÓGICA

La variable Contexto solo presenta el valor «Artístico-humanístico» en la prueba de evaluación P-MU, por lo que no es necesario el cálculo de los estadísticos para esta prueba debido a que la relación estadística es del 100,00%.

Teniendo esto en cuenta, se encuentran las siguientes relaciones entre variables:

a) Las variables Contexto y Demanda cognitiva presentan una relación estadísticamente significativa (0,01) para el estadístico V de Cramer cuando se analiza el conjunto de las pruebas.
b) La variable Contexto mantiene una relación estadísticamente significativa con la variable Tipo de respuesta en el conjunto del total de las pruebas analizadas (0,00), con un valor de 0,282 y en la prueba P-CM (0,03) con un valor de 0,523.
c) La variable Contexto mantiene una relación estadísticamente significativa (0,04) para el conjunto de las pruebas analizadas, con la variable Capacidad competencial, siendo 0,448 el valor del estadístico V de Cramer.
d) La variable Contexto mantiene una relación estadísticamente significativa con la variable Contenido en la prueba P-CM (0,22) y con el conjunto de las pruebas (0,00), siendo el valor del estadístico 0,508 y 0,264, respectivamente.
e) La variable Tipo de respuesta mantiene una relación estadísticamente significativa con la variable Contenido en el conjunto de las pruebas (0,01), siendo 0,345 el valor del estadístico.

Todas estas relaciones no se dan en el conjunto de la muestra (m = 169) y en cada uno de los subconjuntos analizados (P-ME, P-CM, P-MU, P-GA y P-NA). Sin embargo, al analizar las variables Subcompetencia y Demanda cognitiva se obtienen los siguientes resultados:

f) Estas variables presentan una relación directa estadísticamente significativa en todas las pruebas analizadas y en el conjunto de estas (tabla 6), aunque la relación de dependencia entre variables no es muy alta.

Tabla 6.
Valores de V de Cramer para las variables Demanda cognitiva y Subcompetencia

Prueba de examen			*Valor*	*Aprox. Sig.*
P-ME	Nominal por Nominal	V de Cramer	0,622	0,001
	N.º de casos válidos 40			
P-MU	Nominal por Nominal	V de Cramer	0,539	0,018
	N.º de casos válidos 37			
P-CM	Nominal por Nominal	V de Cramer	0,775	0,000
	N.º de casos válidos 25			
P-NA	Nominal por Nominal	V de Cramer	0,768	0,000
	N.º de casos válidos 31			
P-GA	Nominal por Nominal	V de Cramer	0,429	0,039
	N.º de casos válidos 36			
Total	Nominal por Nominal	V de Cramer	0,554	0,000
	N.º de casos válidos 169			

g) El estadístico Lambda de Goodman y Kruskal nos permite establecer la reducción del error al predecir entre variables:
   a) En la prueba P-NA, si las variables son simétricas, existe una probabilidad de reducción del error en la predicción del 17,4 %, con una significación del 0,03.
   b) En el conjunto de las pruebas, si consideramos las variables simétricas, permiten una reducción en el error de predicción del 12,9 % de los casos con una significación del 0,00; en el caso de que la variable Demanda cognitiva sea la variable dependiente, se establece un porcentaje de reducción del error en la predicción del 10,5 %, con una significación del 0,00.

h) El estadístico Tau de Goodman y Kruskal también nos permite establecer porcentualmente la reducción del error al predecir las variables, pero en este caso se basa en las probabilidades de asignación especificadas por las proporciones marginales o condicionales. La tabla 7 muestra los resultados obtenidos para este estadístico, solo se muestran los estadísticamente significativos para simplificar la información.

Tabla 7.
Valores de Tau de Goodman y Kruskal
para las variables Demanda cognitiva y Subcompetencia

Prueba de examen	*Variable dependiente*	*Reducción del error (%)*	*Aprox. sig.*
P-ME	Demanda cognitiva	10,4	0,027
	Subcompetencia	35,5	0,002
P-MU	Subcompetencia	44,5	0,000
P-CM	Demanda cognitiva	17,8	0,046
	Subcompetencia	39,7	0,004
P-NA	Demanda cognitiva	11,1	0,018
	Subcompetencia	59,0	0,001
P-GA	Subcompetencia	26,6	0,005
Total	Demanda cognitiva	5,2	0,000
	Subcompetencia	27,4	0,000

a) Este estadístico nos muestra que las variables analizadas son predecibles en las pruebas y en el conjunto de estas, excepto para las pruebas P-MU y P-GA, en las que no se puede predecir la variable demanda cognitiva cuando es dependiente.
b) Si la variable dependiente es la subcompetencia, la capacidad de reducción en el error en la predicción de las variables es mayor en todos los conjuntos estudiados, por lo que es más probable predecir la demanda cognitiva a partir de la subcompetencia.

CONCLUSIÓN

En esta investigación se marca como primer objetivo comprobar la adecuación de las pruebas analizadas a sus marcos de referencia.

La variable Contexto no se adecúa a la prueba modelo en ninguna de las pruebas analizadas, y presenta mayor diversidad de contextos en la prueba P-ME; a nuestro parecer esta es una fortaleza que debiera generalizarse, ya que al plantear mayor diversidad de contextos se conseguirá reducir la brecha entre los contextos de la prueba y los de la vida cotidiana (OCDE, 2017; MECD, 2014 y Kuglitsch, 2015).

Ninguna de las pruebas analizadas se adecúa a la prueba modelo o al MG en lo referente a la variable Contenido. Esto puede deberse a la diversidad de desarrollos curriculares de las comunidades autónomas. La normativa española establece unos contenidos mínimos que se deben alcanzar al final de la etapa educativa de Educación Primaria y cada comunidad autónoma los amplía y adapta a sus singularidades. En este sentido, el MG sería algo que tener en cuenta en términos generales, y de ahí la variabilidad en los contenidos presentes en cada prueba.

En la variable Formato de pregunta todas las pruebas analizadas se ajustan a lo establecido por el MG y presentan distribuciones de los valores diferentes al de la prueba modelo. Sin embargo, debemos señalar que no compartimos las distribuciones preestablecidas desde el MG, ya que consideramos necesaria la argumentación para el desarrollo de la competencia científico-tecnológica (Alexaindre, 2011; Pinar y Gülüzar, 2017), así como para su evaluación. Si bien es cierto que para la evaluación de la competencia científica este tipo de pruebas deberían complementarse con otros instrumentos de evaluación, como portafolios o diarios de observación, también lo es que en las pruebas escritas se pueden plantear preguntas para que el alumnado argumente, proporcionando rúbricas para que los evaluadores/correctores de estas pruebas evalúen las respuestas a estas.

La variable Demanda cognitiva no se ajusta a lo establecido en el MG debido a que presenta en mayor medida las demandas de bajo nivel. Este hecho se viene corroborando en diversas investigaciones y programas de evaluación externa (Gallardo-Gil, Mayorga-Fernández y Sierra-Nieto, 2014; Yus *et al.*, 2013) y pone en duda el valor de este tipo de pruebas para la evaluación de la competencia científico-tecnológica.

En definitiva, podemos dar como respuesta al primer objetivo que las pruebas analizadas no se ajustan a los marcos de referencia establecidos, a excepción de la variable Tipo de respuesta, que, por otra parte, es la más sencilla de controlar.

En respuesta al segundo objetivo, podemos concluir que las pruebas analizadas consideran la competencia científico-tecnológica de manera sesgada, al tener menor presencia las subcompetencias «Evaluar y diseñar investigaciones científicas» e «Interpretar datos científicamente». Por el contrario, sería recomendable una distribución más equilibrada de los valores de la variable Subcompetencia para evaluar adecuadamente la competencia científica y tecnológica.

El tercer objetivo buscaba determinar si existen relaciones entre las variables analizadas. En este sentido, se debe remarcar que la correlación entre la variable Contexto y el resto de variables es significativa y se encuentra en valores cercanos o superiores al establecido para las Ciencias Sociales (0,3), lo

que nos permite afirmar que, en las pruebas analizadas, cuando se modifica el contexto de un ítem se modifican el resto de variables de estudio. Además, la relación estadística entre las variables Demanda cognitiva y Subcompetencia es significativa y fuerte, e incluso en algunos casos como P-ME, P-CM y P-NA se puede hablar de una relación intensa (superior al 0,6).

Otra cuestión analizada es la predictibilidad de las pruebas. Los resultados muestran que las variables son predecibles entre sí y, por tanto, se podrían establecer patrones de asociación entre estas. Según los datos obtenidos, al seleccionar un ítem tendremos reducciones superiores al 10 % en el error esperado al predecir valores de estas variables si no sabemos el valor de ninguna de ellas, valores entre el 5 y el 11 % si conocemos el valor de la demanda cognitiva y entre el 27,4 y el 59 % si conocemos la subcompetencia. En definitiva, se concluye que en las pruebas analizadas las demandas cognitivas más básicas se asocian a las subcompetencias de menor nivel, y viceversa, y podremos predecirlas reduciendo el margen de error.

Para finalizar, a modo de resumen, para la mejora de las futuras pruebas se podría sugerir utilizar una gran cantidad de contextos, incluir preguntas que requieran argumentación, exigir demandas cognitivas de alto nivel y tener presentes por igual todas las subcompetencias asociadas a la competencia científica.

En estudios posteriores, de similares características, se pretende explorar si las tendencias presentadas en este artículo se consolidan o no en el tiempo.

AGRADECIMIENTOS

Se agradece a los grupos de investigación de excelencia de la Junta de Andalucía, «Didáctica de las Ciencias Experimentales y de la Sostenibilidad» (HUM 613) y «Didáctica de la Matemática. Pensamiento Numérico» (FQM193) por la contribución a esta investigación. El artículo ha sido apoyado por el proyecto EDU2015-70565-P financiado por el Ministerio de Economía y Competitividad de España. Al Ministerio de Educación, Cultura y Deporte, por la concesión de la ayuda FPU15/04972 en la que se enmarca la investigación.

Los comentarios y sugerencias de dos revisores anónimos han mejorado apreciablemente la versión original del presente trabajo.

REFERENCIAS BIBLIOGRÁFICAS

Acevedo, J. A. (2005). TIMSS y PISA. Dos proyectos internacionales de evaluación del aprendizaje escolar en Ciencias. *Revista Eureka sobre Enseñanza y Divulgación de las Ciencias*, *2*(3), 282-301. https://doi.org/10.25267/rev_eureka_ensen_divulg_cienc.2005.v2.i3.01

Álvarez, P. (2016, December 16). La educación española se estanca en Ciencias y Matemáticas y mejora levemente en lectura. *El País*. Recuperado de: http://www.webcitation.org/73jSwvuqO

Anagnostopoulou, K., Hatzinikita, V., Christidou, V. y Dimopoulos K. (2013). PISA Test Items and School-Based Examinations in Greece: Exploring the relationship between global and local assessment discourses. *International Journal of Science Education*, *35*(4), 636-662. https://doi.org/10.25267/rev_eureka_ensen_divulg_cienc.2005.v2.i3.01

Babaci-Wilhite, Z. (Ed.) (2016). *Human rights in language and STEM education: science, technology, engineering and mathematics*. Boston, MA: Sense Publishers.

Benavides, M., León, J. y Etesse, M. (2014). *Desigualdades educativas y segregación en el sistema educativo peruano. Una mirada comparativa de las pruebas PISA 2000 y 2009*. Lima: Grupo de Análisis para el Desarrollo (GRADE).

Berube, C. T. (2014). *STEM and the city: a report on STEM education in the great American urban public school system*. Charlotte, NC: Information Age Publishing.

Black, P. y Wiliam, D. (2018). Classroom assessment and pedagogy. *Assessment in Education: Principles, Policy y Practice*, 1-25.
https://doi.org/10.1080/0969594x.2018.1441807

Bybee, R. W. (1997). *Achieving scientific literacy: From purposes to practices*. Portsmouth, NH: Heinemann.

Caraballo, R., Rico, L. y Lupiáñez, J. L. (2013). Cambios conceptuales en el marco teórico competencial de PISA: el caso de las Matemáticas. *Profesorado. Revista de Curriculum y Formación de Profesorado*, *17*(2), 225-241.

Coughlan, S. (2016, December 6). Pisa tests: Singapore top in global education rankings. BBC. Recuperado de: http://www.webcitation.org/73jTJnUyO

Couso, D. (2014). De la moda de «aprender indagando» a la indagación para modelizar: una reflexión crítica. *XXVI Encuentro de Didáctica de las Ciencias Experimentales*. Huelva (Andalucía).

Custodio Fitó, E., Márquez, C. y Sanmartí, N. (2015). Aprender a justificar científicamente a partir del estudio del origen de los seres vivos. *Enseñanza de las Ciencias. Revista de Investigación y Experiencias Didácticas*, *33*(2), 133.
https://doi.org/10.5565/rev/ensciencias.1316

Feinstein, N. W. (2011). Salvaging science literacy. *Science Education*, *95*(1), 168-185.
https://doi.org/10.1002/sce.20414

Fensham, P. J. (1985). Science for all: A reflective essay. *Journal of Curriculum Studies*, *17*, 415-435.
https://doi.org/10.1080/0022027850170407

Furió, C., Vilches, A., Guisasola, J. y Romo, V. (2001). Finalidades de la enseñanza de las Ciencias en la Secundaria Obligatoria. ¿Alfabetización científica o preparación propedéutica? *Enseñanza de las Ciencias. Revista de Investigación y Experiencias Didácticas*, *19*(3), 365-376.

Gallardo-Gil, M., Fernández-Navas, M., Sepúlveda-Ruiz, M.-P., Serván, M., Yus, R. y Barquín, J. (2010). PISA y la Competencia Científica: Un análisis de las pruebas de PISA en el área de ciencia. *Relieve*, *16*(2), 1-17.
https://doi.org/10.7203/relieve.16.2.4138

Gallardo-Gil, M., Mayorga-Fernández, M. J. y Sierra-Nieto, J. E. (2014). La competencia de «conocimiento e interacción con el mundo físico y natural»: Análisis de las pruebas de evaluación de diagnóstico de Andalucía. *Revista Eureka sobre Enseñanza y Divulgación de las Ciencias*, *11*(2), 160-180.
https://doi.org/10.25267/rev_eureka_ensen_divulg_cienc.2014.v11.i2.04

Goodman, L. A. y Kruskal, W. H. (1963). Measures of Association for Cross Classifications III: Approximate Sampling Theory. *Journal of the American Statistical Association*, *58*(302), 310-364.
https://doi.org/10.1007/978-1-4612-9995-0_3

Hackling, M. W., Ramserger, J. y Chen, H. L. S. (2017). *Quality Teaching in Primary Science Education*. Basel: Springer.
https://doi.org/10.1007/978-3-319-44383-6

Hatzinikita, V., Dimopoulos, K. y Christidou, V. (2008). PISA test items and school textbooks related to science: a textual comparison. *Science Education*, *92*, 664-687.
https://doi.org/10.1002/sce.20256

Hernández, M. I., Couso, D. y Pintó, R. (2015). Analyzing students' learning progressions throughout a teaching sequence on Acoustic Properties of Materials with a model-based inquiry approach. *Journal of Science Education and Technology*, *24*(2), 356-377.
https://doi.org/10.1007/s10956-014-9503-y

Hodson, D. y Reid, D. J. (1988). Science for all: motives, meanings and implications. *School Science Review*, *69*, 653-661.

Jiménez-Aleixandre, M. P. (2011). 10 ideas clave. Competencias en argumentación y uso de pruebas. *Educatio Siglo XXI*, *29*(1), 363-366.

Jerrim, J., López-Agudo, L. A., Marcenaro-Gutiérrez, O. D. y Shure, N. (2017). What happens when econometrics and psychometrics collide? An example using the pisa data. *Economics of Educational Review*, *61*(1), 51-58.
https://doi.org/10.1016/j.econedurev.2017.09.007

Kuglitsch, R. (2015). Teaching for Transfer: Reconciling the Framework with Disciplinary Information Literacy. *Libraries and the Academy*, *15*(3), 457-470.
https://doi.org/10.1353/pla.2015.0040

Kwak, Y. (2017). Analysis of Features of Korean Fourth Grade Students' timss Science Achievement in Content Domains with Curriculum Change. *Journal of the Korean Association for in Science Education*, *37*(4), 599-609.

Lau, K. C. (2009). A critical examination of pisa's assessment on scientific literacy. *International Journal of Science and Mathematics Education*, *7*, 1061-1088.
https://doi.org/10.1007/s10763-009-9154-2

Lee, O. (2017). Common Core State Standards for ela/Literacy and Next Generation Science Standards: Convergences and Discrepancies Using Argument as an Example. *Educational Researcher*, *46*(2), 90-102.
https://doi.org/10.3102/0013189x17699172

Leguilloux, C. (2018, February 22). L'Education nationale est-elle incapable de s'évaluer. Boursier. Recuperado de: http://www.webcitation.org/73jTSJEwF

Martin, A., Gómez, R. y Verde, E. (2017). Primary and Secondary Physical Education: comparative analysis about teacher and students interaction. *Profesorado. Revista de Curriculum y Formacion de Profesorado*, *21*(2), 253-270.

Martin, M. O., Mullis, I. V. S., Foy, P. y Hooper, M. (2015). *timss 2015 International Results in Science*. Boston: International Association for the Evaluation.

Mayring, P. (2017). Evidence triangulation in health research The combination of experimental, descriptive and content-analytical approaches. *Kolner Zeitschrift Fur Soziologie und Sozialpsychologie*, *69*(2), 415-434.

MECD (2014). Real Decreto 126/2014, de 28 de febrero, por el que se establece el currículo básico de la Educación Primaria (2014). Recuperado de: https://www.boe.es/buscar/pdf/2014/BOE-A-2014-2222-consolidado.pdf

MECD (2015a). Marco General de la evaluación final de Educación Primaria. Recuperado de: http://www.mecd.gob.es/dctm/inee/evaluacionfinalprimaria/marco-teorico-evaluacion-final-6ep.pdf?documentId=0901e72b81ceacce

MECD (2015b). Orden ECD/65/2015, de 21 de enero, por la que se describen las relaciones entre las competencias, los contenidos y los criterios de evaluación de la educación primaria, la educación secundaria obligatoria y el bachillerato. Recuperado de: https://www.boe.es/diario_boe/txt.php?id=BOE-A-2015-738

MECD (2015c). Prueba de la competencia en ciencia y Tecnología. Recuperado de: http://www.mecd.gob.es/dctm/inee/evaluacionsextoprimaria/pruebamodelo6epcyt.pdf?documentId=0901e72b81d3c84f

Mihladiz, G. y Dogan, A. (2017). Investigation of the Pre-service Science Teachers' Pedagogical Content Knowledge about the Nature of Science. *Hacettepe Universitesi Egitim Fakultesi Dergisi-Hacettepe University Journal of Education*, *32*(2), 380-395.

OCDE (2017). PISA 2015 Science Framework. *En PISA 2015 Assessment and Analytical Framework: Science, Reading, Mathematics, Financial Literacy and Collaborative Problem Solving*. OECD Publishing: Paris.
https://doi.org/10.1787/9789264281820-3-en

OCDE (2018). What do science teachers find most satisfying about their work? *PISA in Focus*, *81*(1), 1-7.
https://doi.org/10.1787/3e711a23-en

Ouariachi, T., Olvera-Lobo, M. D. y Gutiérrez-Pérez, J. (2017). Evaluation of online games for teaching and learning on climate change. *Enseñanza de las Ciencias. Revista de Investigación y Experiencias Didácticas*, *35*(1), 193.
https://doi.org/10.5565/rev/ensciencias.2088

Pholphirul, P. (2017). Pre-primary education and long-term education performance: Evidence from Programme for International Student Assessment (PISA) Thailand. *Journal of Early Childhood Reseach*, *15*(4), 410-432.
https://doi.org/10.1177/1476718x15616834

Pinar, S. C. y Gülüzar, E. (2017). Developing Students' Scientific Writing and Presentation Skills through Argument Driven Inquiry: An Exploratory Study. *Journal of Chemical Education*, *94*, 837-843.
https://doi.org/10.1021/acs.jchemed.6b00915

Roberts, D. A. (2007). *Scientific literacy/science literacy.* En S. K. Abell y N. G. Lederman (Eds.), *Handbook of research on science education* (pp. 729-780). Mahwah: Lawrence Erlbaum Associates.

Sampson, V., Grooms, J. y Walker, J. P. (2011). Argument-Driven Inquiry as a way to help students learn how to participate in scientific argumentation and craft written arguments: An exploratory study. *Science Education*, *95*(2), 217-257.
https://doi.org/10.1002/sce.20421

Sanmartí, N. (2007). *Evaluar para aprender: 10 ideas claves*. Barceló: Graó.

Santos, A. (2016, December 22). CC. OO. y UGT rechazan el informe de la Comunidad sobre el bilingüismo. *El País*. Recuperado de: http://www.webcitation.org/73jTqGLHD

Sjøberg, S. (2015). Foreword. En E. K. Henriksen, J. Dillon y J. Ryder (Eds.), *Understanding student participation and choice in Science and Techonology education* (pp. 5-7). Dordrecht: Springer.

Sjöström, J., Eilks, I. y Zuin, V. G. (2016). Towards eco-reflexive science education. *Science y Education*, *25*(3), 321-341.
https://doi.org/10.1007/s11191-016-9818-6

Solano-Flores, G., Contreras-Niño, L. A. y Backhoff-Escudero, E. (2006). Traducción y adaptación de pruebas: Lecciones aprendidas y recomendaciones para países participantes en TIMSS, PISA y otras comparaciones internacionales. *Revista Electrónica de Investigación Educativa*, 8(2), 1-21.

Stanley, G., MacCann, R., Gardner, J., Reynolds, L. y Wild, I. (2009). *Review of teacher assessment: what works best and issues for development*. Oxford University Centre for Educational Development.

Teixeira, E. S. y Greca, I. M. (2015). La enseñanza de la gravitación universal de Newton orientada por la historia y la filosofía de la ciencia: una propuesta didáctica con un enfoque en la argumentación. *Enseñanza de las Ciencias. Revista de Investigación y Experiencias Didácticas*, 1, 205-223.
https://doi.org/10.5565/rev/ensciencias.1226

Vidal, R., Díaz, M. A. y Jarquín, H. (2004). *Resultados de las pruebas PISA 2000 y 2003 en México: habilidades para la vida en estudiantes de 15 años*. Madrid: INEE.

Yus Ramos, R., Fernández Navas, M., Gallardo Gil, M., Barquín Ruiz, J., Sepúlveda Ruiz, M. P. y Serván Núñez, M. J. (2013). La competencia científica y su evaluación. Análisis de las pruebas estandarizadas de PISA. *Revista de Educación*, 360, 557-576.
http://doi.org/10.4438/1988-592X-RE-2011-360-127

Analysis of the external assessment of the scientific-technological competence in 6th grade of Primary Education (2016)

Tobías Martín-Páez
Dpto. de Didáctica de las Ciencias Experimentales,
Universidad de Granada, Granada, España
tmartin@ugr.es

Javier Carrillo-Rosúa
Dpto. de Didáctica de las Ciencias Experimentales,
Universidad de Granada, Granada, España
Instituto Andaluz de Ciencias de la Tierra (CSIC-UGR),
Armilla, España
fjcarril@ugr.es

José Luís Lupiáñez-Gómez
Dpto. de Didáctica de las Matemáticas,
Universidad de Granada, Granada, España
lupi@ugr.es

José Miguel Vílchez-González
Dpto. de Didáctica de las Ciencias Experimentales,
Universidad de Granada, Granad, España
jmvilchez@ugr.es

Scientific and technological competence is the educational answer to a modern view on scientific literacy. For the assessment of this competence, educational institutions have tried to respond with many proposals, thus becoming a subject of growing social interest with frequent references in the media (e. g. Álvarez, 2016; Coughlan, 2016; Santos, 2016; Leguilloux, 2018).

An option adopted by several educational organizations and institutions is the design and application of external tests to assess the quality of educational systems by evaluating the development of student competences. Some examples of these external tests are the Programme for International Student Assessment (PISA) (OCDE, 2018) or the Trends in International Mathematics and Science Study (TIMSS) (Martin, Mullis, Foy and Hooper, 2015).

But these tests are analyzed through research works focused on discussing the imperfections in the elaboration of the items or inaccuracies in their theoretical frameworks. Solano-Flores, Contreras-Niño and Backhoff-Escudero (2006), for example, study the linguistic limitations of the PISA tests when they are written in English and have to be translated into the different languages of the target countries. The weaknesses of their theoretical frameworks have been analyzed in publications such as Acevedo (2005), in the scientific field, and Caraballo, Rico and Lupiáñez (2013), in mathematics.

In this article we present a documentary analysis that studies the available five external tests of evaluation of the scientific and technological competence elaborated for the sixth year of Primary Education in the 2015-2016 academic year and different Autonomous Communities in Spain. This study aims: *a*) to prove the adjustment and adaptation of the test to their reference theoretical frameworks; *b*) to characterize the presence of scientific and technological competence and how it is considered in the tests; *c*) to develop a model that can determine the relationships between these variables and the evaluated competence.

Five variables drawn from the General Framework of the Final Assessment of Primary Education from the Spanish Ministry of Education, Science and Sports (contexts, contents, cognitive demand and question format) and from the 2015 PISA Framework of the OECD (sub-competences) are considered in this study. A total of 169 activities have been analyzed through triangulation among experts and researchers in Science Education. The obtained data were analyzed with SPSS, and V-Cramer, Lambda, Tau-Godman, Kruskal and descriptive statistics were calculated.

The results related to context, content and cognitive demands show weakness in the adjustment to the theoretical framework. Among its causes, the diversity of regional curricula and the traditional approach to content are cited. Moreover, there is adjustment to the theoretical framework in all the tests with regards to the question format; but the model is criticized for not measuring relevant skills such as argumentation.

One can also observe that, in the test, the scientific competence is biased, as it is measured according to today's frameworks of competence such as PISA. Thus, more activities should be focused on data interpretation and research design.

Finally, relationships between the studied variables have been found: *a*) context shows a close connection with other variables; *b*) cognitive demand and sub-competence is significative, and in some test, strong. These relations determine predictability of the questions.

To conclude, some suggestions for future tests are thereof derived: *a*) to use a large number of contexts, including questions that require argumentation and high-level cognitive demands, and *b*) to bear in mind all the sub-competences associated with the scientific competence.