

Diseño y construcción de un instrumento de evaluación de la competencia matemática: aplicabilidad práctica de un juicio de expertos

Ramón García Perales ^a

Resumo

Se realiza una aproximación a una de las etapas fundamentales en el diseño y construcción de cualquier instrumento de evaluación educativa, la valoración realizada por un grupo de expertos en la materia objeto de la evaluación. La tabla de especificaciones se conforma como la herramienta idónea para desarrollar un juicio de expertos lo que facilita, entre otros objetivos, calcular el índice de validez de contenido de cada uno de sus elementos. A modo de ejemplo, se explica el proceso desarrollado para facilitar la construcción de un instrumento que evalúa competencia matemática. Tras contar con un diseño inicial de 64 ítems, se han eliminado 11 ítems y se han entresacado múltiples valoraciones cualitativas que han permitido la mejora significativa del instrumento, gracias a la participación de 51 expertos en el campo de las matemáticas. Para este instrumento se ha obtenido una validez de contenido elevada, siendo la mayoría de los índices calculados superiores a .80.

Palabras clave: Instrumento de evaluación. Juicio de expertos. Tabla de especificaciones. Validez de contenido. Competencia matemática.

1 Introducción

La definición operativa del constructo a medir constituye el primer paso para la elaboración de un instrumento de evaluación. Con este proceso se pretende delimitar qué es lo que se va a evaluar, por lo que habrá que ser conscientes de que “es necesaria una precisa definición del constructo a medir para asegurar la validez de contenido” (CASTRO MORERA, 2011, p. 119). En este proceso de definición, el juicio de expertos en el área objeto de evaluación cobra una especial

^a Universidad de Castilla La Mancha. Albacete, Espanha.

Recebido em: 26 jan. 2017
Aceito em: 27 out. 2017

importancia, utilizando la tabla de especificaciones como recurso y tomando el cálculo de la validez de contenido como finalidad.

Cabero-Almenara y Llorente-Cejudo (2013) consideran que la técnica del juicio de expertos “consiste, básicamente, en solicitar a una serie de personas la demanda de un juicio hacia un objeto, un instrumento, un material de enseñanza, o su opinión respecto a un aspecto concreto” (p. 14). Esta técnica constituye un indicador prioritario para calcular el índice de validez de contenido, requiriendo rigurosidad estadística y metodológica para que el instrumento valorado pueda ser utilizado para la finalidad para la cual fue diseñado.

Según Escobar-Pérez y Cuervo-Martínez (2008), “el concepto esencial de validez de contenido es que los ítems de un instrumento de medición deben ser relevantes y representativos del constructo para un propósito evaluativo particular” (p. 27). Estos investigadores definen el juicio de expertos como “una opinión informada de personas con trayectoria en el tema, que son reconocidas por otros como expertos cualificados en éste, y que pueden dar información, evidencia, juicios y valoraciones” (2008, p. 29).

Esta evaluación cobra una enorme importancia ya que colaboran en la eliminación de aquellos ítems irrelevantes y en el ajuste de aquellos otros que lo requieran para su mantenimiento, todo con la finalidad de valorar la pertinencia, coherencia y calidad del instrumento. Con ello, en este artículo se realiza una aproximación a su conceptualización y se propone un ejemplo de procedimiento para su realización, en este caso para la competencia matemática.

En la actualidad, el desarrollo de evaluaciones por competencias constituye una prioridad para organismos nacionales e internacionales. Por medio de ellas se realiza una aproximación al concepto de competencia y a la forma de evaluarlas. En esta evaluación por competencias resulta imprescindible utilizar instrumentos de valoración criterial cuyo contenido guarde una relación estrecha con lo que los alumnos aprenden en los centros educativos. En este sentido, en la delimitación del marco conceptual de la competencia matemática se ha realizado un acercamiento al *Programme for International Student Assessment* (PISA) por ser un referente fundamental en la evaluación por competencias, especialmente las ediciones de 2003 y 2012 por ser las evaluaciones que valoran específicamente la competencia matemática de los alumnos de 15 años, aunque también se señalan aspectos de la edición de 2015 por ser la más reciente aunque no tomase la evaluación de la competencia matemática como prioritaria.

Esta competencia se define como la habilidad para manejar y utilizar los números y sus operaciones básicas, los símbolos y las formas de expresión y razonamiento matemático, con la finalidad de interpretar la realidad y así poder resolver problemas relacionados con la vida cotidiana y el mundo laboral (GARCÍA PERALES, 2014). PISA 2012 y 2015 la definen como:

Capacidad del individuo para formular, emplear e interpretar las matemáticas en distintos contextos. Incluye el razonamiento matemático y la utilización de conceptos, procedimientos, datos y herramientas matemáticas para describir, explicar y predecir fenómenos. Ayuda a los individuos a reconocer el papel que las matemáticas desempeñan en el mundo y a emitir los juicios y las decisiones bien fundadas que los ciudadanos constructivos, comprometidos y reflexivos necesitan (ORGANIZATION FOR ECONOMIC COOPERATION AND DEVELOPMENT, 2016, p. 74).

Para establecer qué se va a valorar a través de un instrumento, primeramente se deberá de clarificar la conceptualización del constructo propio a evaluar, es decir, determinar y especificar sus estándares de ejecución y sus respectivos indicadores de evaluación. A partir de estos estándares se delimitan una serie de dimensiones cuyo desarrollo permite la clarificación del constructo tomado como punto de partida. En última instancia, se tendrá que realizar un análisis de los componentes de aprendizaje de cada dimensión para operativizar y especificar los indicadores para evaluarlas de manera particular y el constructo a nivel general. El desarrollo de estos indicadores pretende orientar y facilitar una evaluación de calidad, válida y fiable, que favorezca la identificación de diferentes niveles de competencia dentro del constructo. Jornet et al. (2011) manifiestan que:

Las competencias, como constructos complejos, suelen implicar diversos tipos de subdominios o subcompetencias (habilidades o destrezas, actitudes, valores, etc.), por lo que el proceso de planificación y diseño de cualquier competencia partirá de un análisis cuidadoso de los subdominios competenciales que se integran en cada una de ellas (p. 127–128).

De acuerdo a PISA 2012 y 2015 (INSTITUTO NACIONAL DE EVALUACIÓN EDUCATIVA, 2013; ORGANIZATION FOR ECONOMIC COOPERATION AND DEVELOPMENT, 2016), la estructura de la evaluación de la competencia matemática viene definida en función de los siguientes aspectos:

- Contenido matemático. Engloba cuatro dominios de conocimiento: cantidad, espacio y forma, cambio y relaciones e incertidumbre y datos.
- Procesos matemáticos. Incluye los procesos matemáticos de formular, emplear e interpretar/valorar y las capacidades matemáticas subyacentes denominadas: razonar y argumentar, matematizar, elaborar estrategias para resolver problemas, representar, comunicar, usar lenguaje formal, técnico, simbólico y las operaciones y, por último, usar herramientas matemáticas. Estas acciones cognitivas son repartidas en tres grupos de competencia según PISA 2012 (no reseñadas en PISA 2015): reproducción (aplicación del conocimiento estudiado), conexión (apoyada en las anteriores y son situaciones de resolución de problemas alejadas de la rutina aunque incluyen situaciones familiares para el alumnado) y reflexión (análisis de los procesos necesarios y/o empleados para resolver un problema).
- Situaciones. Denominadas en PISA 2015 como contextos aunque la conceptualización es similar a PISA 2012. Abarcan aquellos contextos próximos al mundo real en el que el alumno se desenvuelve y en los que se utilizan las matemáticas. Pueden ser: personales (contexto inmediato del alumno: familia, grupo de iguales y propio alumno), laborales (entorno laboral o escuela, en PISA 2015 esta categoría es denominada profesional aunque la conceptualización es similar), sociales (relacionadas con el contexto social) y científicas (propio campo de las matemáticas y análisis de los procesos relacionados con la tecnología).
- Una vez realizada la conceptualización de la competencia matemática prestando una especial atención a PISA, se realiza un acercamiento al proceso de diseño y construcción de un instrumento de evaluación para esta competencia que toma como base la fundamentación matemática especificada anteriormente. En el desarrollo de este proceso resulta imprescindible poner en marcha un juicio de expertos, es decir, contar la valoración de personas especialistas en el campo objeto de la evaluación, la competencia matemática, aspecto que constituye la finalidad principal de la presente investigación.

2 Método

Como ha quedado patente, el objetivo de este artículo es hacer una aproximación a cómo diseñar y construir un instrumento de evaluación, en sus primeras fases de elaboración, señalando un ejemplo concreto para la competencia matemática. Previamente al desarrollo del juicio de expertos, se ha tenido que hacer una aproximación terminológica y operacional del constructo a evaluar para fijar las dimensiones y los indicadores que han servido de guía para la definición de los ítems.

El diseño y construcción de la Batería de Evaluación de la Competencia Matemática (BECOMA) constituye el aspecto central de este artículo, mostrándose el juicio de expertos realizado para su configuración. Este instrumento ha sido elaborado por García Perales (2014) y muestra elevados índices de fiabilidad y validez. Se trata de una batería conformada por cuatro subcompetencias que constituyen el contenido de la competencia matemática escolar: Estadística y probabilidad, Aritmética, Geometría y Magnitudes y proporcionalidad. Corresponden a los dominios de conocimiento definidos en PISA 2012 y 2015 como Incertidumbre y datos, Cantidad, Espacio y forma y Cambio y relaciones respectivamente.

En la BECOMA estas subcompetencias aparecen representadas mediante 64 ítems repartidos en ocho pruebas de evaluación. Los ítems establecidos guardan una relación estrecha con los contenidos y criterios de evaluación del área de Matemáticas y con los indicadores de evaluación de la competencia matemática, tomando como referencia los marcos normativos educativos vigentes. Estas ocho pruebas quedan repartidas en el diseño inicial entre las cuatro subcompetencias señaladas anteriormente, quedando su configuración tal y como se muestra en la Tabla 1:

En las Tablas 2 y 3 puede observarse el número de ítems para cada prueba y subcompetencia, mostrándose los porcentajes que representan.

Tabla 1. Configuración de la batería de evaluación de la competencia matemática (BECOMA).

Subcompetencias	Pruebas
Estadística y probabilidad	Prueba nº 1 Interpretación matemática
Aritmética	Pruebas nº 2 Cálculo mental, nº 4 Series lógicas numéricas y nº 5 Descubriendo algoritmos
Geometría	Pruebas nº 3 Propiedades geométricas y nº 7 Series lógicas de figuras
Magnitudes y proporcionalidad	Pruebas nº 6 Unidades convencionales y nº 8 Invención de problemas

Fuente: Elaboración propia (2014).

Tabla 2. Porcentaje y número de ítems para cada subcompetencia.

Subcompetencia	Peso en % en el total de la batería	Total de ítems	Nº de prueba	Nº de ítems
Estadística y probabilidad	16	10	1	10
			2	8
Aritmética	34	22	4	10
			5	4
Geometría	22	14	3	7
			7	7
Magnitudes y proporcionalidad	28	18	6	10
			8	8
Total	100	64	8	64

Fuente: Elaboración propia (2014).

Tabla 3. Número de ítems dedicado a cada subcompetencia y porcentaje que representa.

Prueba	Estadística y probabilidad	Aritmética	Geometría	Magnitudes y proporcionalidad
1	10	-	-	-
2	-	8	-	-
3	-	-	7	-
4	-	10	-	-
5	-	4	-	-
6	-	-	-	10
7	-	-	7	-
8	-	-	-	8
Total (64 ítems)	10 (16%)	22 (34%)	14 (22%)	18 (28%)

Fuente: Elaboración propia (2014).

En las evaluaciones PISA, los ítems que engloban los ejercicios no solo aparecen aglutinados según el dominio de conocimiento o área de contenido en la que se incluyen. También son agrupados en función de las actividades cognitivas que movilizan (procesos) y los contextos en los que se encuadran las tareas presentadas (situaciones), observándose en la BECOMA tal conceptualización (Tablas 4 y 5), tomando en consideración la base teórica de PISA 2012 por evaluar de manera prioritaria la competencia matemática.

Tabla 4. Número de ítems de la batería para cada proceso.

Procesos o actividades cognitivas	Número de ítems en la batería
Grupo de reproducción	15 (ítems pruebas 2 y 3)
Grupo de conexión	37 (ítems pruebas 1, 4, 6 y 7)
Grupo de reflexión	12 (ítems pruebas 5 y 8)

Fuente: Elaboración propia (2014)

Tabla 5. Número de ítems de la batería para cada situación.

Situaciones	Número de ítems en la batería
Personales	8 (ítems prueba 2)
Laborales	24 (ítems pruebas 3, 4 y 7)
Sociales	20 (ítems pruebas 1 y 6)
Científicas	12 (ítems pruebas 5 y 8)

Fuente: Elaboración propia (2014).

Una vez realizado el diseño inicial del instrumento se ha procedido a hacer un proceso de depuración y mejora de esta versión inicial para contar con una primera selección de ítems. Esto se hace mediante la colaboración de un grupo de especialistas en el área de Matemáticas, más adelante se muestran las características distintivas de estos expertos, con arreglo a su ajuste al marco de la evaluación definido y demandando un equilibrio entre cada una de las subcompetencias, sus contenidos e indicadores. Esta recogida de información permite un conocimiento exhaustivo del funcionamiento del instrumento con la finalidad de dar la mayor objetividad posible a su estructura interna y a sus resultados. Este análisis ofrece una información valiosa y constituye el primer paso en la validación.

El principal instrumento para desarrollar un juicio de expertos es la tabla de especificaciones. Se trata de una herramienta destinada a comprobar en qué grado una prueba mide realmente lo que se desea medir. Debido a ello, su objetivo es observar la validez de contenido del instrumento, es decir, comprobar si los ítems que conforman cada una de las ocho pruebas recogen información válida según los objetivos y las subcompetencias establecidas. De esta manera facilita la comprobación sobre si los elementos integrantes en

el instrumento resultan adecuados y suficientes o, por el contrario, aparecen ítems sin incluir o han existido otros cuyo contenido no aportan información de interés para los objetivos fijados. Según Bardín (1986) las cualidades y requisitos necesarios para satisfacer las demandas definidas con el diseño de un instrumento son: exhaustividad, exclusión mutua, homogeneidad, objetividad, pertinencia y productividad.

Otros aspectos que se han estudiado con la tabla de especificaciones son: analizar si el peso otorgado a cada bloque de ítems en el diseño inicial del instrumento se ajusta al considerado por los expertos y valorar la redacción y el grado de ajuste de los ítems y de las instrucciones facilitadas de acuerdo a la edad y al nivel de los alumnos.

Teniendo todo esto presente, en el proceso de depuración de la BECOMA se han utilizado los siguientes instrumentos:

- Una carta de presentación en la que se explica qué se pretende con la investigación y qué se les demanda como especialistas en el área de Matemáticas. Así, incluía propósito, objetivos generales y específicos y sistema de variables e indicadores.
- La batería diseñada inicialmente.
- La tabla de especificaciones compuesta por las siguientes partes:
 - Presentación e instrucciones. Se realiza una breve introducción agradeciendo su colaboración y se explican las cuatro subcompetencias a evaluar dentro de la competencia matemática, cómo se definen y cómo se evaluarán a través de las distintas pruebas.
 - Valoración del porcentaje otorgado a cada subcompetencia. Inicialmente se ha adjudicado un peso a cada subcompetencia en función del número de ítems utilizados para su evaluación. Los expertos deben valorar qué peso darían a cada una de ellas en el conjunto de un instrumento de evaluación de la competencia matemática.
 - Valoración de la relevancia de cada uno de los ítems (Figura 1). Se les solicita que valoren si cada uno de los ítems es adecuado para la consecución del objetivo perseguido con cada una de las pruebas de la batería.

De acuerdo con cada uno de los objetivos y utilizando las pruebas elaboradas, indique con una “X” el grado de relevancia de cada ítem para la **CONSECUCCIÓN DEL OBJETIVO DE LA PRUEBA**, siendo la escala a utilizar la siguiente:

“**Sí**” = considero el ítem **esencial** para la consecución del objetivo de la prueba

“**No**” = considero el ítem **innecesario** para la consecución del objetivo de la prueba

“**?**” = no tengo claro si el ítem es **esencial o innecesario**

Estadística y Probabilidad

Objetivo	Prueba	Ítem	Sí	No	?
"Precisar, elegir y aplicar las estrategias adecuadas para resolver problemas de estadística y probabilidad"	1	1			
		2			
		3			
		4			
		5			
		6			
		7			
		8			
		9			
		10			

Fuente: Elaboración propia (2014).

Figura 1. Valoración de la relevancia de los ítems para la subcompetencia Estadística y Probabilidad.

A partir de las puntuaciones se calcula el Índice de Validez de Contenido (IVC). Se define como el grado en el que el instrumento presenta una muestra adecuada de los contenidos a los que hace referencia, sin omisiones y sin desequilibrios (KERLINGER; LEE, 2002). Lawshe (1975) establece la fórmula para calcularla: $IVC = [ne - (N/2)] / (N/2)$, donde ne es el número de expertos que han valorado el ítem como esencial y N el número total de expertos que han evaluado el ítem. El IVC oscila entre +1 y -1, siendo las puntuaciones positivas las que indican una mayor validez de contenido. Cuanto

más cercano el índice a +1 mejor es la validez. Lawshe sugiere que un IVC de .29 será adecuado cuando hayan participado al menos 40 expertos en la valoración.

Valoración de la redacción y del grado de adaptación de cada ítem (Figura 2). Se les ha pedido que valorasen cada ítem de forma independiente según estos aspectos: presentación, redacción, claridad y ajuste.

A continuación vamos a valorar la **REDACCIÓN DE CADA ÍTEM**. Utilizando la batería elaborada se va a evaluar el grado en el que la presentación es adecuada, los ítems están bien redactados, si queda claro lo que al alumno se le pide y si los ítems están ajustados para la evaluación de los alumnos de 5°. De esta forma, para cada ítem, marque con una “X” siguiendo la escala siguiente:

“**Sí**” = considero **adecuada** la redacción de cada ítem

“**No**” = considero **inadecuada** la redacción de cada ítem

“**?**” = no tengo claro si la redacción de cada ítem es **adecuada o inadecuada**

Estadística y Probabilidad

Prueba 1	Ítem	Sí	No	?	Observaciones
La presentación es adecuada. La redacción es correcta. Lo que se le pide al alumno está claro. Lo que se le requiere al alumno está ajustado para el nivel de 5°.	1				
	2				
	3				
	4				
	5				
	6				
	7				
	8				
	9				
	10				

Fuente: Elaboración propia (2014).

Figura 2. Valoración de la redacción de los ítems para la subcompetencia Estadística y Probabilidad

Valoración de las instrucciones generales de la batería y de las específicas de cada prueba (Figura 3). Se ha buscado averiguar si las instrucciones facilitadas eran comprensibles y ajustadas a lo que se pedía al alumno en las pruebas. El instrumento está conformado para ser aplicado en dos sesiones, contando cada una de ellas con unas instrucciones propias. Se ha buscado también contar con su opinión sobre la portada de la batería.

Por último, siguiendo los indicadores de evaluación anteriores se va a realizar una valoración cualitativa de la portada, las instrucciones generales y las establecidas para cada prueba. Para ello debe de contestar en función de:

Por último, siguiendo los indicadores de evaluación anteriores se va a realizar una valoración cualitativa de la portada, las instrucciones generales y las establecidas para cada prueba. Para ello debe de contestar en función de:

“**Sí**” = considero **adecuada** su presentación

“**No**” = considero **inadecuada** su presentación

“**?**” = no tengo claro si su presentación es **adecuada o inadecuada**

	Campo	Sí	No	?	Observaciones
La presentación es adecuada. La redacción es correcta. Lo que se le pide al alumno está claro. Lo que se le requiere al alumno está ajustado para el nivel de 5º. Las instrucciones son fácilmente entendibles.	Portada				
	Instrucciones Generales				
	Prueba 1				
	Prueba 2				
	Prueba 3				
	Prueba 4				

Fuente: Elaboración propia (2014).

Figura 3. Valoración de las instrucciones generales de la batería y de las específicas de las cuatro primeras pruebas.

La tabla de especificaciones ha sido cumplimentada por 51 expertos docentes repartidos entre maestros de Educación Primaria (total de 39, de los que 23 de ellos estaban impartiendo Matemáticas en 5º de Educación Primaria, nivel al que va dirigida la batería, el resto trabajaban en otros niveles de esta etapa educativa), docentes de Matemáticas de Educación Secundaria (10) y profesorado de Enseñanza Universitaria de dicha especialidad en Facultades de Educación (2). Esta discrepancia en las cifras de selección de docentes a favor de los de Educación Primaria ha sido debida a que el instrumento se inserta en esta etapa educativa, buscando así asegurar su funcionalidad, representatividad y eficacia y, por ende, su éxito futuro en la administración. La selección de los participantes se ha desarrollado de la siguiente manera:

- Docentes de Educación Primaria: aprovechar los colegios que iban a participar en el proceso de investigación ya que la batería se administraría a sus alumnos de 5º.
- Docentes de Educación Secundaria: contacto por teléfono y por correo electrónico con equipos directivos y jefes de departamento de Matemáticas de distintos institutos.
- Docentes de Enseñanza Universitaria: contacto por correo electrónico y organización de entrevistas personales para mostrarles la finalidad de la investigación y solicitar su colaboración.

En la selección se han elegido personas de la zona geográfica en la que se ha desarrollado la investigación (provincia de Albacete, España), con una experiencia docente mínima de diez años, con una especial cercanía, disponibilidad y motivación por participar en la investigación, imparcialidad a la hora de emitir juicios, confianza en sí mismos y adaptabilidad y, por último, gusto e interés por la enseñanza de las Matemáticas. Para su selección se ha realizado una entrevista previa con cada uno de ellos tomando en consideración criterios de selección como los señalados (SKJONG; WENTWORHT, 2001). Además se les ha informado de la finalidad de la investigación y de lo que se esperaba del proceso de validación que se iba a desarrollar. En cuanto a la cantidad de jueces, Gable y Wolf (1993), Grant y Davis (1997) y Lynn (1986) sugieren un rango entre dos y veinte, mientras que Hyrkäs, Appelqvist-Schmidlechner y Oksa (2003) afirman que con diez sería suficiente.

3 Resultados

Los resultados se indican en función de lo requerido mediante cada uno de los elementos de la tabla de especificaciones:

1. Valoración del porcentaje otorgado a cada subcompetencia.
2. Valoración de la relevancia de cada uno de los ítems según el objetivo de la prueba en la que se incluyen.
3. Valoración de la redacción y del grado de adaptación de cada ítem.
4. Valoración de las instrucciones generales de la batería y de las específicas de cada prueba.

3.1 Valoración del porcentaje otorgado a cada subcompetencia

Uno de los aspectos analizados mediante la tabla de especificaciones ha sido el peso de cada subcompetencia en el total de la batería. Con ello se ha pretendido saber si el porcentaje otorgado a cada una de ellas a partir del número de ítems que las representa ha sido similar al peso que cada experto les dedicaría en el desarrollo de un instrumento de evaluación que midiera el mismo constructo, es decir, la competencia matemática. Para favorecer la visualización de las aportaciones se han fijado unos intervalos para ubicar cada puntuación dada.

En el diseño inicial de la BECOMA el porcentaje total otorgado a cada subcompetencia ha sido el siguiente: Estadística y probabilidad 16%, Aritmética 34%, Geometría 22% y Magnitudes y proporcionalidad 28%. En la Tabla 6 se pueden ver las aportaciones realizadas, observándose que el porcentaje inicial de tres de las cuatro subcompetencias se ha incluido dentro del intervalo más representado por los expertos.

Al comparar las aportaciones de los expertos con las evaluaciones PISA 2003 y 2012 (INEE, 2008, 2013) por centrarse de manera prioritaria en la evaluación de la competencia matemática y PISA 2015 (OECD, 2016) por ser la última evaluación desarrollada, el peso de cada subcompetencia o dominio de conocimiento fue el siguiente:

3.2 Valoración de la relevancia de cada uno de los ítems según el objetivo de la prueba en la que se incluyen

Uno de los objetivos fundamentales de la cumplimentación de la tabla de especificaciones es examinar la contribución de los ítems a la consecución del objetivo de la prueba en la que se han incluido. Para ello se ha calculado el Índice de Validez de Contenido (IVC) de los ítems, las subcompetencias en las que

se incluyen y el global del instrumento utilizando la media aritmética. En esta valoración se han utilizado las opciones Sí-No-?, escogiendo solo las respuestas afirmativas para el cálculo de este índice.

En la Tabla 8 se especifican los resultados, observándose altos índices de validez de contenido para la casi totalidad de los ítems. A pesar de ello, dentro de las subcompetencias Aritmética, Geometría y Magnitudes y proporcionalidad, hay ítems con un índice por debajo de .25 (ítems 18, 23, 30 y 42), por lo que se eliminarán de la nueva versión de la batería. Según las observaciones realizadas, las causas de

Tabla 6. Intervalos y número de expertos ubicados en cada uno de ellos.

Intervalo (%)	Estadística y probabilidad	Aritmética	Geometría	Magnitudes y proporcionalidad
0-5	0	0	0	0
6-10	6	0	0	0
11-15	16	0	7	3
16-20	19	0	13	11
21-25	9	5	14	16
26-30	1	8	10	15
31-35	0	16	6	6
36-40	0	11	1	0
41-45	0	7	0	0
46-50	0	4	0	0
51 ó más	0	0	0	0
Total de expertos	51	51	51	51

Fuente: Elaboración propia (2014).

Tabla 7. Porcentaje adjudicado según distintas fuentes y momentos de la investigación y por las evaluaciones PISA.

Subcompetencia / Dominio de conocimiento en PISA	Batería Inicial	Juicio de expertos	PISA 2003	PISA 2012	PISA 2015
Estadística y probabilidad/ Incertidumbre	16	16-20	24	25	25
Aritmética/Cantidad	34	31-35	27	25	25
Geometría/Espacio y forma	22	21-25	24	25	25
Magnitudes y proporcionalidad/ Cambios y relaciones	28	21-25	25	25	25

Fuente: Elaboración propia (2017).

Tabla 8. Índice de validez de contenido (IVC) de los ítems.

Subcompetencia	Ítem	No	?	Sí	IVC
Estadística y probabilidad	1	0	1	50	.96
	2	5	5	41	.61
	3	1	4	46	.80
	4	5	1	45	.76
	5	0	3	48	.88
	6	4	0	47	.84
	7	2	3	46	.80
	8	0	0	51	1
	9	3	1	47	.84
	10	2	3	46	.80
	11	1	1	49	.92
	12	1	3	47	.84
	13	2	4	45	.76
	14	0	0	51	1
	15	1	1	49	.92
	16	2	1	48	.88
	17	2	2	47	.84
	18	14	10	27	.06
Aritmética	26	3	0	48	.88
	27	1	1	49	.92
	28	2	2	47	.84
	29	1	0	50	.96
	30	14	6	31	.22
	31	6	1	44	.73
	32	0	2	49	.92
	33	1	3	47	.84
	34	2	6	43	.69
	35	1	8	42	.65
	36	3	1	47	.84
	37	3	2	46	.80
	38	4	1	46	.80
	39	3	3	45	.76

Continúa

Continuación

	19	2	1	48	.88
	20	2	0	49	.92
	21	0	1	50	.96
	22	2	4	45	.76
	23	7	12	32	.25
	24	4	3	44	.73
	25	1	3	47	.84
Geometría	50	3	1	47	.84
	51	1	2	48	.88
	52	6	1	44	.73
	53	10	2	39	.53
	54	1	1	49	.92
	55	4	2	45	.76
	56	3	2	46	.80
	40	0	1	50	.96
	41	1	0	50	.96
	42	15	7	29	.14
	43	1	2	48	.88
	44	1	3	47	.84
	45	0	2	49	.92
	46	1	1	49	.92
	47	1	0	50	.96
Magnitudes y proporcionalidad	48	10	1	40	.57
	49	1	1	49	.92
	57	0	1	50	.96
	58	1	1	49	.92
	59	1	1	49	.92
	60	2	1	48	.88
	61	1	2	48	.88
	62	1	3	47	.84
	63	2	2	47	.84
	64	2	3	46	.80

Fuente: Elaboración propia (2014).

un índice tan bajo han sido las siguientes: ítems 18 y 30 no es posible realizar el algoritmo y los ítems 23 y 42 no son un contenido específico de la subcompetencia en la que se incluyen. Los ítems 8 y 14 han mostrado una validez de contenido perfecta.

Los índices globales reflejados en la siguiente Tabla 9 reflejan una adecuada validez de toda la batería y de sus subcompetencias. Polit y Hungler (2000) afirman que cuando este índice es mayor o igual a .80 es indicativo de validez de contenido elevada, lo que ha ocurrido en el 73,44% de los componentes de la batería. En esta línea Voutilainen y Liukkonen (1995) afirman que un índice de .80 sería suficiente para mantener un ítem en el diseño del instrumento.

Por último, el cálculo de la concordancia entre jueces ha sido fundamental, es decir, conocer el grado de acuerdo entre ellos en las estimaciones realizadas. La aplicación del estadístico Kappa ha dado un índice de .82, reflejando un grado de acuerdo elevado entre los evaluadores.

Tabla 9. Índice de Validez de Contenido (IVC) para cada subcompetencia y para el total de la batería.

Subcompetencia	IVC
Estadística y probabilidad	.83
Aritmética	.78
Geometría	.77
Magnitudes y proporcionalidad	.84
Total	.81

Fuente: Elaboración propia (2014).

Tabla 10. Relación entre los ítems suprimidos y los que permanecen.

Suprimido		Permanece	
Ítem	IVC	Ítem	IVC
10	.80	9	.84
17	.84	16	.88
35	.65	33 y 34	.84 y .69
39	.76	36, 37 y 38	.84, .80 y .80
53	.53	54 y 56	.92 y .80
63	.84	57, 59 y 61	.96, .92 y .88
64	.80	58, 60 y 62	.92, .88 y .84

Fuente: Elaboración propia (2014).

3.3 Valoración de la redacción y del grado de adaptación de cada ítem

Se trata de una valoración cualitativa de la redacción de los ítems y su ajuste al nivel y edad de los alumnos. Para ello se han utilizado las opciones Sí-No-? para la valoración de este grado de adecuación en función de los indicadores establecidos. El análisis cualitativo ha permitido mejorar los ítems, cada prueba y la batería en su conjunto.

En líneas generales han valorado positivamente la mayoría de los ítems. Entre las aportaciones realizadas, tres de ellas han permitido la eliminación de 7 ítems: 10, 17, 35, 39, 53, 63 y 64. Se trata de ítems que valoraban el mismo contenido, se presentaban al alumno de una manera similar o bien requerían para su solución de procesos de ejecución parecidos. Los ítems seleccionados para ser eliminados han sido los que tenían un índice de validez de contenido menor tal y como refleja la Tabla 10, a pesar de que el de algunos de ellos se ha encontrado por encima de .80:

3.4 Valoración de las instrucciones generales de la batería y de las específicas de cada prueba

Al igual que en el apartado anterior, se ha realizado una evaluación cualitativa con las opciones Sí-No-? para la valoración de la portada del instrumento y las instrucciones de cada una de las dos sesiones y de cada una de las ocho pruebas que componen la BECOMA.

Esta valoración y las observaciones realizadas han permitido avanzar hacia el ajuste y mejora del instrumento. A modo de ejemplo se pueden señalar aportaciones tales como: añadir el nombre y número de lista de cada alumno en la portada, poner un ejemplo en la prueba 3, quitar dos minutos al tiempo de aplicación de la prueba 5, dejar más espacio entre problemas en la prueba 8, etc.

El juicio de expertos ha permitido la reducción y mejora de los ítems del instrumento, con la consecuente bajada de su tiempo de aplicación tal y como se muestra en las siguientes Tablas 11 y 12.

4 Discusión

El nacimiento del concepto de competencia constituye uno de los grandes cambios producidos en la educación contemporánea, siendo los distintos elementos constitutivos del currículo los referentes que permiten su desarrollo y adquisición.

Tabla 11. Comparativa de los ítems de la batería en el diseño inicial y tras realizar el juicio de expertos.

Prueba/Denominación	Batería Inicial	Juicio de expertos
Nº 1 Interpretación matemática	10	9
Nº 2 Cálculo mental	8	6
Nº 3 Propiedades geométricas	7	6
Nº 4 Series lógicas numéricas	10	8
Nº 5 Descubriendo algoritmos	4	3
Nº 6 Unidades convencionales	10	9
Nº 7 Series lógicas de figuras	7	6
Nº 8 Inventión de problemas	8	6
Total	64	53

Fuente: Elaboración propia (2014).

Tabla 12. Comparativa de la temporalización de la batería en el diseño inicial y tras realizar el juicio de expertos.

Prueba/Denominación	Batería Inicial	Juicio de expertos
Nº 1 Interpretación matemática	17	16
Nº 2 Cálculo mental	4	4,5
Nº 3 Propiedades geométricas	15	14
Nº 4 Series lógicas numéricas	10	8
Nº 5 Descubriendo algoritmos	8	6
Nº 6 Unidades convencionales	16	15
Nº 7 Series lógicas de figuras	7	6
Nº 8 Inventión de problemas	16	12
Total	93	81,5

* Tiempo en minutos

Fuente: Elaboración propia (2014).

Un alumno es considerado “competente” cuando es capaz de movilizar recursos y estrategias para resolver situaciones de la vida real. Los contenidos dejan de ser el eje vertebrador del currículo y los contextos en los que se desarrolla el aprendizaje son fundamentales.

La integración de este término en el ámbito educativo emana básicamente de los trabajos que ha venido desarrollando la OECD desde finales de los 90 al poner en marcha el Proyecto DeSeCo de definición y selección de competencias en el ámbito educativo (OECD, 2005). La forma establecida por la OECD para evaluar estas competencias es a través del proyecto PISA, utilizándose sus resultados para diseñar e implementar reformas educativas que permitan mejorar los logros educativos de los escolares (MESA, GÓMEZ Y CHEAH, 2013).

En este artículo se ha generalizado la fundamentación teórica de PISA al diseño de un instrumento que mide competencia matemática. Jornet Meliá y González Such (2009) inciden en la necesidad de evaluaciones criterioles que rompan con la inadecuación de los sistemas normativos de construcción de test psicométricos para la elaboración de pruebas estandarizadas de rendimiento. Esto viene a justificar tanto el diseño de la BECOMA, parte de los aprendizajes matemáticos básicos incluidos en el currículum escolar, como la participación de expertos que ejercen como docentes en activo en centros educativos.

La evaluación criterial de la competencia matemática implica decidir cuáles son las matemáticas que hacen a una persona competente en este campo, siendo imprescindible establecer unos descriptores breves, concretos y ubicados en contextos específicos (PLAZA MENÉNDEZ, 2013). La OECD (2016) remarca esta importancia al manifestar que la competencia matemática “contribuye a que los individuos sean conscientes del papel que desempeñan las matemáticas en el mundo y les ayuda a emitir los juicios y las decisiones bien fundadas que se exigen a los ciudadanos constructivos, comprometidos y reflexivos” (p. 74).

Tal y como se ha explicado, el marco conceptual de la BECOMA parte de lo que los docentes enseñan por medio de sus programaciones didácticas y su currículum correspondiente. Por ello la utilización del juicio de expertos forma parte de la estimación de la validez de contenido, ya que contar con sus valoraciones y aportaciones tiene una especial importancia para adaptar e enriquecer un instrumento de evaluación diseñado. En este artículo se ha explicado la conceptualización y el desarrollo de este proceso, aplicándolo en la práctica a un instrumento de evaluación que pretende medir competencia matemática.

Este proceso ha sido desarrollado por 51 docentes con una trayectoria profesional mínima de diez años de experiencia docente.

Sus aportaciones y sugerencias han permitido la adecuación, relación y congruencia de los ítems con la variable que se pretende medir y el rediseño del instrumento de medición en pro de la investigación, es decir, que “mide aquello que realmente pretende medir o sirve para el propósito para el que ha sido construido” (MARTÍN ARRIBAS, 2004, p. 27). Esto ha facilitado después someterlo a la confiabilidad por medio de varias administraciones a diferentes grupos de alumnos y el cálculo de distintos estadísticos, obteniéndose índices de fiabilidad y validez de constructo entre .73 y .90.

El diseño inicial de la BECOMA estaba conformado por 64 ítems con un tiempo de aplicación de 93 minutos. A partir de una primera versión del instrumento se ha calculado el índice de validez de contenido, gracias a la cumplimentación de una tabla de especificaciones. A partir de esta valoración, los ítems con un índice de validez de contenido por debajo de .25 han sido eliminados, 4 en total. En general la validez de contenido del instrumento ha sido elevada, ya que la mayoría de los índices han sido superiores a .80 (POLIT y HUNGLER, 2000), siendo el índice global de la batería de .81. Tras el cálculo de la concordancia en las estimaciones entre jueces se ha observado un elevado consenso entre ellos a la hora de asignar las puntuaciones, ya que el estadístico Kappa ha dado un índice de .82

Junto al índice de validez de contenido, en la tabla de especificaciones se han considerado otros aspectos importantes para el diseño y la construcción del instrumento. En primer lugar se ha puesto en relación el peso dedicado inicialmente a cada subcompetencia con el valor que los expertos darían a cada una de ellas en un instrumento que evalúe competencia matemática, observándose que los porcentajes dados a tres de las cuatro subcompetencias se han incluido dentro del intervalo fijado por los expertos. Únicamente no ha coincidido en la subcompetencia Magnitudes y proporcionalidad.

Otro aspecto valorado ha sido la evaluación cualitativa de la redacción de cada ítem y el grado de ajuste al nivel y edad de los alumnos. Esto ha permitido la eliminación de 7 ítems. Por último han realizado una valoración de las instrucciones generales de la batería y de las específicas de cada prueba, lo que ha facilitado ir en la línea de construir un instrumento cada vez más válido y fiable.

La selección exhaustiva de expertos en Matemáticas mediante una entrevista previa personal ha permitido la viabilidad de todo el proceso desarrollado.

Además se ha buscado evitar la ambigüedad en las tareas a realizar junto a una forma de presentación atractiva y motivadora, con vistas a reducir errores y aumentar la objetividad de las aportaciones obtenidas. El incluir columnas de observaciones ha permitido contar con una mayor cantidad de información sobre aspectos no evaluados de manera cuantitativa. Todo ello ha facilitado el diseño y construcción de un instrumento suficiente y representativo de acuerdo a las finalidades propuestas en la investigación y los marcos legales y curriculares existentes para el área de Matemáticas. Aún así, hemos de ser conscientes de que existen aspectos que han resultado difíciles de controlar, como ocurre en cualquier juicio de expertos, como es el nivel de dificultad de la tarea. La revisión del proceso desarrollado facilitará su mejoramiento continuado (ESCOBAR-PÉREZ y CUERVO-MARTÍNEZ, 2008).

En definitiva este instrumento ha sido diseñado para que favorezca la evaluación formativa y dinámica del alumnado teniendo presentes los contextos, los conocimientos y los aprendizajes a alcanzar. En el logro de esta finalidad, el desarrollar un juicio de expertos y calcular el índice de validez de contenido constituyen el inicio de un posible éxito posterior, es decir, la construcción de un instrumento válido y fiable. Esto ha quedado patente a lo largo de este artículo, por lo que su desarrollo constituye un aspecto central en la construcción de instrumentos de evaluación educativa.

Desenho e construção de um instrumento de avaliação da competição matemática: aplicabilidade prática de um julgamento de especialistas

Resumo

Uma aproximação é feita para uma das etapas fundamentais no desenho e construção de qualquer instrumento de avaliação educacional, a avaliação feita por um grupo de especialistas no assunto para avaliar. A tabela de especificações é uma ferramenta ideal para desenvolver um julgamento especializado, o que facilita, entre outros objetivos, o cálculo do índice de validade de conteúdo de cada um de seus elementos. Como exemplo, é explicado o processo desenvolvido para facilitar a construção de um instrumento que avalia a competência matemática. Depois de ter um desenho inicial de 64 itens, 11 itens foram eliminados e várias estimativas qualitativas foram obtidas, o que permitiu uma melhoria significativa do instrumento, graças à participação de 51 especialistas no campo da matemática. Para este instrumento, foi obtida uma alta validade de conteúdo, com a maioria dos índices calculados superiores a .80.

Palavras-chave: Instrumento de avaliação. Julgamento de especialistas. Tabela de especificações. Validade de conteúdo. Competência matemática.

Design and construction of an evaluation instrument of mathematical competence: Practical applicability of an expert assessment

Abstract

We approach one of the fundamental stages in the design and construction of any instrument of educational evaluation: the assessment carried out by a group of experts in the subject matter. The chart of specifications emerges as the suitable tool to develop an expert assessment that enables us, among other objectives, to estimate the index of content validity of each of the elements. For example, the developed process is explained to facilitate the construction of an instrument that evaluates mathematical competence. After an initial design of 64 items, 11 items were eliminated and multiple qualitative valuations were addressed to allow for the significant improvement of the instrument, after to the participation of 51 experts in the field of the mathematics. We obtained high content validity for the instrument, and most estimated indexes were higher than 80.

Keywords: Evaluation instrument, Expert judgment, Chart of specifications, Content validity, Mathematical competence.

Referencias

BARDÍN, L. *Análisis de contenido*. Madrid: Akal, 1986.

CABERO-ALMENARA, J.; LLORENTE-CEJUDO, M. C. La aplicación del juicio de experto como técnica de evaluación de las tecnologías de la información (TIC). *Eduweb: Revista de Tecnología de Información y Comunicación en Educación*, v. 7, n. 2, p. 11–22, 2013. Disponível em: <<http://servicio.bc.uc.edu.ve/educacion/eduweb/index.html>>. Acesso em: 9 dez. 2016.

CASTRO MORERA, M. ¿Qué sabemos de la medida de las competencias? Características y problemas psicométricos en la evaluación de competencias. *Bordón: Revista de Pedagogía*, v. 63, n. 1, p. 109–123, 2011.

ESCOBAR-PÉREZ, J.; CUERVO-MARTÍNEZ, A. Validez de contenido y juicio de expertos: una aproximación a su utilización. *Avances en Medición*, v. 6, n. 1, páginas 27–36, 2008. Disponível em: <http://www.humanas.unal.edu.co/psicometria/files/7113/8574/5708/Articulo3_Juicio_de_expertos_27-36.pdf>. Acesso em: 9 jan. 2017.

GABLE, R. K.; WOLF, M. B. *Instrument development in the affective domain: Measuring attitudes and values in corporate and school settings*. Boston: Kluwer Academic Publishers, 1993.

GARCÍA PERALES, R. *Diseño y validación de un instrumento de evaluación de la competencia matemática: rendimiento matemático de los alumnos más capaces*. Tesis (Doctorado en Educación). Facultad de Educación – UNED, Madrid, 2014. Disponível em: <<http://e-spacio.uned.es/fez/view/tesisuned:Educacion-Rgarcia>>. Acesso em 11 dez. 2017.

GRANT, J. S.; DAVIS, L. L. Selection and use of content experts for instrument development. *Research in Nursing & Health*, v. 20, n. 3, p. 269–274, 1997. doi: [https://doi.org/10.1002/\(SICI\)1098-240X\(199706\)20:3<269::AID-NUR9>3.0.CO;2-G](https://doi.org/10.1002/(SICI)1098-240X(199706)20:3<269::AID-NUR9>3.0.CO;2-G)

HYRKÄS, K.; APPELQVIST-SCHMIDLECHNER, K.; OKSA, L. Validating an instrument for clinical supervision using an expert panel. *International Journal of Nursing Studies*, v. 40, n. 6, p. 619–625, 2003. doi: [http://dx.doi.org/10.1016/S0020-7489\(03\)00036-1](http://dx.doi.org/10.1016/S0020-7489(03)00036-1)

INSTITUTO NACIONAL DE EVALUACIÓN EDUCATIVA (INEE). *PISA 2003 Matemáticas: Informe español*. Madrid: Ministerio de Educación, Cultura y Deporte, 2008.

INSTITUTO NACIONAL DE EVALUACIÓN EDUCATIVA (INEE). *PISA 2012 Informe Español*. Madrid: Ministerio de Educación, Cultura y Deporte, 2013. v. I: Resultados y contexto.

JORNET MELIÁ, J. M.; GONZÁLEZ SUCH, J. Evaluación criterial: determinación de estándares de interpretación (EE) para pruebas de rendimiento educativo. *Estudios sobre Educación*, n. 16, p. 103–123, 2009. Disponível em: <http://www.uv.es/gem/gemhistorico/publicaciones/Evaluacion_criterial_Determinacion_de_estandares_de_interpretacion_EE_para_pruebas_de_rendimiento_educativo.pdf>. Acesso em: 14 jan. 2017.

JORNET MELIÁ, J. M. et al. Diseño de procesos de evaluación de competencias: consideraciones acerca de los estándares en el dominio de las competencias. *Bordón: Revista de Pedagogía*, v. 63, n. 1, p. 125–145, 2011.

KERLINGER, F. N.; LEE, H. B. *Investigación del comportamiento: métodos de investigación en ciencias sociales*. Ciudad de México: McGraw-Hill, 2002.

LAWSHE, C. H. A quantitative approach to content validity. *Personnel Psychology*, v. 28, n. 4 p. 563–575, 1975. doi: 10.1111/j.1744-6570.1975.tb01393.x

LYNN, M. R. Determination and quantification of content validity. *Nursing Research*, v. 35, n. 6, p. 382–385, 1986.

MARTÍN ARRIBAS, M. C. Diseño y validación de cuestionarios. *Matronas Profesión*, v. 5, n. 17, p. 23–29, 2004. Disponível em: <http://www.enferpro.com/documentos/validacion_cuestionarios.pdf>. Acesso em: 4 jan. 2017.

MESA, V. M.; GÓMEZ, P.; CHEAH, U. H. Influence of international studies of student achievement on mathematics teaching and learning. In: CLEMENTS, M. A. K. et al. (Ed.). *Third international handbook of mathematics education*. Dordrecht: Kluwer, 2013. p. 861–900.

ORGANIZATION FOR ECONOMIC COOPERATION AND DEVELOPMENT. *The definition and selection of key competencies: Executive summary*, 2005. Disponível em: <<https://www.oecd.org/pisa/35070367.pdf>>. Acesso em: 14 jan. 2017.

ORGANIZATION FOR ECONOMIC COOPERATION AND DEVELOPMENT. *PISA 2015 Assessment and analytical framework: Science, reading, mathematic and financial literacy*. Paris: OECD Publishing, 2016.

PLAZA MENÉNDEZ, P. Las competencias matemáticas en el aprendizaje a lo largo de la vida. *Suma*, n. 72, páginas 9–15, 2013.

POLIT D. F.; HUNGLER B. P. *Investigación científica en Ciencias de la Salud: principios y métodos*. Ciudad de México: McGraw-Hill Interamericana, 2000.

SKJONG, R.; WENTWORTH, B. H. Expert Judgement and risk perception. In: INTERNATIONAL OFFSHORE AND POLAR ENGINEERING CONFERENCE STAVANGER, 2001. *Anais...*, Norway: [s.n.], 2001.

VOUTILAINEN, P.; LIUKKONEN, A. Senior Monitor: laadun arviointi- mittarin sisällön validiteetin määrittäminen. *Hoitotiede*, v. 7, n. 1, p. 51–56, 1995.



Informações do autor

Ramón García Perales: Doctor en Educación. Profesor del Departamento de Pedagogía de la Facultad de Educación de Albacete, Universidad de Castilla La Mancha (UCLM). Investigador de la Facultad de Educación (UCLM). Contacto: ramon.garciaperales@uclm.es