

CoDiLe: un instrumento para evaluar el conocimiento disciplinar de lengua española de los maestros en formación

CoDiLe: An instrument for evaluating the Spanish-language disciplinary knowledge of pre-service teachers

María-Teresa MUEDRA-PERIS. Estudiante de Doctorado. Universidad de Valencia (muepe@alumni.uv.es).

Dr. Manuel MONFORT-PAÑEGO. Profesor. Universidad de Valencia (manuel.monfort@uv.es).

Dra. Ángela GÓMEZ-LÓPEZ. Profesora. Universidad de Valencia (angela.gomez@uv.es).

Dra. Eva MORÓN-OLIVARES. Profesora. Universidad de Valencia (eva.moron@uv.es).

Resumen:

En las últimas décadas, diversas investigaciones se han centrado en evaluar el conocimiento didáctico del profesorado, en especial en ciencias y en matemáticas. Sin embargo, en el área de lengua, pocos trabajos han diseñado y validado herramientas con este fin. El presente trabajo pretende avanzar en esta línea de investigación con el diseño y la validación de un cuestionario para evaluar el conocimiento disciplinar de lengua española de los maestros en formación. Participaron estudiantes españoles de grado de Magisterio y de máster de Formación del Profesorado de Secundaria, así como expertos en didáctica de la lengua y en validación de cuestionarios. Para el análisis de la validez de contenido, se utilizó el método Delphi y, para el

estudio de la consistencia, se aplicó un análisis psicométrico a través del método de fiabilidad test-retest. El instrumento se mostró consistente y válido. Los resultados estuvieron por debajo de lo esperado y desvelaron que la muestra presentaba un claro déficit en contenido disciplinar en lengua española. Estos datos parecen estar en línea con los obtenidos para otras áreas. Por tanto, CoDiLe puede contribuir a definir y subsanar estas posibles deficiencias mediante la aportación de datos consistentes a los formadores de maestros que permitan una orientación más efectiva de sus prácticas.

Palabras clave: instrumento de medida, nivel de conocimientos, lengua española, formación de profesores.

Fecha de recepción del original: 20-07-2023.

Fecha de aprobación: 25-10-2023.

Cómo citar este artículo: Muedra-Peris, M. T., Monfort-Pañeco, M., Gómez-López, Á., y Morón-Olivares, E. (2024). CoDiLe: un instrumento para evaluar el conocimiento disciplinar de lengua española de los maestros en formación [CoDiLe: An instrument for evaluating the Spanish-language disciplinary knowledge of pre-service teachers]. *Revista Española de Pedagogía*, 82 (288), 271-290. <https://doi.org/10.22550/2174-0909.4037>

Abstract:

In recent decades, several studies have focused on assessing teachers' pedagogical knowledge, especially in Science and Mathematics. However, with regards to languages, few studies have designed and validated instruments for this purpose. This paper aims to advance in this line of research: a questionnaire is designed and validated to evaluate pre-service teachers' content knowledge in Spanish language. Spanish undergraduate and master students, as well as experts on the topic and tests, participated in the experiment. For the analysis of content validity, the Delphi method was used. For the study of consistency

over time, a psychometric analysis using the test-retest reliability method was applied. The instrument was found to be consistent and valid. The results on the assessment of students' knowledge were below expectations and revealed that the sample showed a clear deficit in Spanish language content knowledge. These data seem to be in line with those obtained in other studies. Therefore, CoDiLe may contribute to define and address these potential shortcomings and provide teachers with consistent data to improve their practices.

Keywords: measuring instrument, knowledge level, Spanish language, teacher education.

1. Introducción

Evaluar las necesidades educativas de un país implica investigar todos aquellos factores que forman parte del sistema educativo, como la formación del profesorado. Este es un tema de investigación complejo y uno de los ejes sobre los que actuar para mejorar el sistema educativo; no obstante, no hay consenso sobre qué factores promueven la calidad docente y cómo incorporarlos a la formación inicial (Harris y Sass, 2011): algunos estudios se han centrado en el aprendizaje profesional (Opfer y Pedder, 2011), la cognición (Borg, 2003) o el conocimiento personal (Pajares, 1992).

Uno de los autores que mayores aportaciones ha hecho a la investigación sobre la formación del profesorado ha sido Shulman (1987), que propuso un nuevo concepto integrador, el *conocimiento didáctico del contenido* (en adelante, CDC): la

combinación de contenido y pedagogía en la comprensión de cómo se organizan, representan y adaptan ciertos temas a los intereses y a las habilidades de los alumnos, y de cómo se presentan para su enseñanza.

Este nuevo concepto ha actuado como catalizador de importantes investigaciones que, en las últimas décadas, han puesto de manifiesto la diferencia entre el *conocimiento del contenido* (en adelante, CC) y la enseñanza de este (CDC) (Bucat, 2005). Aunque la relación entre el CC y el CDC no está definida con nitidez en la literatura, sí parece claro que el CC está en el centro del desarrollo de las competencias profesionales de los docentes (Kleickman *et al.*, 2013).

En el modelo de Shulman, el CC es el primero de los aspectos que hay que tener en cuenta para el estudio de la enseñan-

za de las disciplinas. Las investigaciones señalan que un CC profundo mejora las explicaciones, favorece el uso de recursos e influye en la comprensión de los estudiantes y en su éxito académico (Chetty *et al.*, 2011); así, resulta imprescindible delimitar qué nivel de conocimientos tienen los futuros profesores para intervenir en la mejora de la formación inicial y permanente (Kleickmann *et al.*, 2013). Conocer el contenido de la materia que se va a impartir es una premisa para poder enseñarla (Friedrichsen *et al.*, 2009). De hecho, en algunas investigaciones, el CC del profesorado explicó de forma significativa la mejora de los resultados de los estudiantes (Gess-Newsome *et al.*, 2019).

En las últimas décadas, diversas investigaciones se han centrado en evaluar el CC del profesorado, en especial en el área de ciencias y matemáticas. Godino *et al.* (2016) evaluaron el conocimiento de los profesores sobre la visualización de objetos tridimensionales en 241 estudiantes de magisterio. Para diseñar el cuestionario, se tomaron investigaciones anteriores, el currículo y libros de texto de gran tirada nacional. A pesar de que las preguntas fueron extraídas de libros de primaria, los resultados obtenidos mostraron que el 62% de los alumnos no contestaron de forma óptima a las tareas propuestas.

El Ministerio de Educación, Cultura y Deporte (2012) participó en un estudio internacional sobre la formación inicial en matemáticas de los maestros. En él, se empleó un cuestionario basado en el estudio de TEDS-M de Tattoo *et al.* (2012), que evaluaba tanto su conocimiento didáctico

como disciplinar. Para diseñar las preguntas, se utilizaron investigaciones anteriores y los marcos legales de los países participantes. El cuestionario final tenía 74 preguntas en formato de elección múltiple o respuesta abierta. Las puntuaciones medias obtenidas por los futuros profesores españoles estaban por debajo de la media internacional, tanto en conocimientos matemáticos como en didácticos, aunque puntuaron algo mejor en estos últimos.

Vásquez y Alsina (2015) validaron un cuestionario de respuesta abierta para evaluar conocimientos didáctico-matemáticos para enseñar probabilidad. Se tomó como base de conocimiento investigaciones anteriores, orientaciones curriculares y libros de texto. Tanto la aplicación piloto del instrumento como su replicación arrojaron resultados medio-bajos en todas las categorías.

Verdugo *et al.* (2019) analizaron el conocimiento didáctico-disciplinar en ciencias de los maestros en formación; para ello, se creó un cuestionario de 30 ítems de alternativa múltiple para los que se tomaron como base el currículo nacional español y los libros de texto españoles. El instrumento mostró un dominio del contenido científico mejorable y la presencia de algunos errores conceptuales relevantes.

En cuanto al CC en lengua, pocos trabajos han diseñado y validado herramientas para evaluarlo; la mayoría de ellos se centran en los conocimientos necesarios para enseñar a leer. Binks-Cantrell *et al.* (2012) y Washburn *et al.* (2016) validaron un instrumento para evaluar el conocimiento

de los profesores sobre los constructos básicos de la lengua que intervienen en la enseñanza de la lectura. Participaron 279 profesores y maestros en formación. El cuestionario incluía 38 ítems orientados al conocimiento del contenido. Los resultados revelaron falta de conocimientos, especialmente en morfología y fonología.

El presente trabajo pretende avanzar en esta línea de investigación: el objetivo propuesto es diseñar y validar un instrumento que permita evaluar el conocimiento del contenido que tienen los maestros en formación en lengua española. Hasta donde se sabe, no existen instrumentos con los que valorar el conocimiento disciplinar que deben tener los docentes (en concreto, los maestros en formación) para impartir lengua española. La necesidad de dicho instrumento es relevante para la investigación en el ámbito educativo por una doble razón: el hecho de que la lengua es, a la vez, contenido disciplinar y vehículo del resto de aprendizajes. Esta necesidad pone en valor el objetivo de este trabajo, pues la creación de instrumentos que permitan evaluar el conocimiento que poseen los docentes sobre las materias del currículo repercute de forma directa sobre los programas de formación del profesorado.

2. Metodología

2.1. Diseño del estudio

Para analizar el CC sobre lengua española de los maestros en formación, se diseñó y validó un cuestionario basado en el método Delphi (Andrés *et al.*, 2019) en cuatro fases:

- Fase 1. Recogida de evidencias. Búsqueda de bibliografía. Selección de los indicadores de evidencia.
- Fase 2. Desarrollo de la versión I. Elaboración de los ítems. Evaluación de expertos.
- Fase 3. Desarrollo de la versión II. Prueba piloto. Evaluación por los estudiantes.
- Fase 4. Desarrollo de la versión final. Primer pase (validez de constructo). Segundo pase (fiabilidad).

Para el estudio de la fiabilidad del cuestionario, se aplicó un test-retest y un análisis psicométrico.

2.2. Participantes

Durante las fases 2 y 3 (validez de contenido), se utilizaron dos grupos de participantes para evaluar el contenido y la comprensibilidad del test inicial. El primer grupo se compuso de seis expertos independientes: tres en filología y didáctica de la lengua, uno en CDC y dos en instrumentos y diseños de investigación (uno de ellos experto también en CDC). La selección de los expertos se hizo de acuerdo con los siguientes criterios: debían estar fuera del estudio, poseer el grado de doctor, ser profesores universitarios y tener publicaciones de calidad en didáctica de la lengua o en CDC, en métodos de investigación o en validación de cuestionarios. En paralelo, participaron como evaluadores dos sujetos externos y no relacionados ni con el contenido ni con los criterios de selección.

El segundo grupo se compuso de un conjunto natural de 53 estudiantes universitarios (19-23 años) de ambos sexos de segundo curso del grado de Maestro en Educación Primaria de una universidad española. Realizaron el test en línea y, con posterioridad, se seleccionó a diez para ser entrevistados. Se les pidió que evaluaran el contenido y la comprensibilidad del test inicial. Los resultados se utilizaron para hacer una estimación previa del funcionamiento del cuestionario.

Para la fase 4, la muestra de participantes fue de 256 estudiantes de ambos sexos (18-25 años) de los grados de Maestro en Educación Primaria/Infantil (cursos primero, segundo y cuarto) de una universidad española. Además, participó un grupo natural de 20 estudiantes (23-35 años) del posgrado del Máster de Profesorado de Educación Secundaria (especialidad Lengua Castellana y su Literatura). Este grupo tenía conocimiento disciplinar más especializado y se utilizó para comprobar si el instrumento discriminaba entre diferentes niveles de conocimiento. De la muestra inicial, 190 estudiantes (152 mujeres y 38 hombres) completaron los dos pases del cuestionario.

2.3. Procedimiento

En la fase 1, se consideraron diversas fuentes para diseñar el cuestionario. Se realizó una búsqueda de literatura especializada sobre conocimiento disciplinar de lengua española y sobre el uso de cuestionarios para su evaluación en la enseñanza del español como lengua materna y extranjera.

En la fase 2, se partió de este análisis para generar un banco de preguntas y se

consensuó una primera versión del cuestionario. Este se envió a través de una plataforma virtual a seis expertos independientes y a dos sujetos externos para su evaluación.

En la fase 3, el instrumento resultante fue sometido a prueba en un grupo de 53 estudiantes para hacer una primera estimación del funcionamiento y de la pertinencia de las preguntas. Para ello, se utilizó una plataforma en línea durante una sesión de clase. El tiempo se limitó a 40 minutos. Con posterioridad, se entrevistó a 10 estudiantes para completar la información obtenida.

Según las sugerencias de ambos grupos, se reformularon o sustituyeron algunas preguntas. La nueva versión se envió a dos expertos, uno en cuestionarios y métodos de investigación y el otro en didáctica de la lengua española, y sus propuestas se incorporaron también en el cuestionario.

Por último, durante la fase 4, se pasó el cuestionario definitivo a la muestra del estudio en dos etapas: con los datos de la primera, se evaluó la validez del constructo y, con ellos y con los de la segunda, se estudió la fiabilidad. Se articularon cuatro condiciones experimentales para contrabalancear el orden de las preguntas en la primera y segunda etapa. Se incluyó una pregunta de control de la atención en la posición 21 (hacia la mitad del cuestionario) en las cuatro condiciones.

Los cuestionarios se administraron a través de la plataforma Moodle web. La participación fue voluntaria. Las instrucciones

estaban escritas al principio del cuestionario y una de las investigadoras también las leyó en voz alta. Se resolvieron las dudas y se limitó el tiempo a 40 minutos. Tras cuatro semanas, se realizó la segunda etapa.

2.4. Análisis de datos

Para estudiar la validez de contenido, se aplicó el modelo Delphi (Mokkink *et al.*, 2010). En la fase 1, las preguntas iniciales se desarrollaron a partir de las categorías validadas por Muedra (2020): morfología, fonética, fonología y ortografía, nivel léxico-semántico, sintaxis, tipología textual, procesos de expresión oral y escrita/pragmática, procesos de comprensión oral y escrita, recursos literarios.

Para concretarlas, se recurrió al último documento puente de la comunidad autónoma en la que se llevó a cabo el estudio, orientado a facilitar la programación de aula (CEFIRE, 2015). Las investigadoras extrajeron indicadores de conocimiento y los clasificaron de forma independiente dentro de cada una de las categorías. Se entiende por indicador de conocimiento una unidad de contenido expresada de forma concreta y objetiva y traducible, en su caso, a una conducta evaluable (Alfaro-Carvajal *et al.*, 2022): por ejemplo, el indicador «El sustantivo. Clases: propios y comunes, individuales y colectivos, concretos y abstractos» se clasificó en la categoría «morfología». En los currículos de índole competencial, la selección de indicadores pretende facilitar la elaboración posterior de los instrumentos y medios de evaluación de las competencias propuestas.

Para evaluar el funcionamiento de estas categorías y de sus indicadores, se clasificaron las actividades de dos colecciones de libros de texto de lengua castellana de primaria. Se observó que estas categorías recogían todos los indicadores, por lo que se utilizaron como referencia para definir las preguntas del cuestionario.

En la fase 2, se diseñó un banco de 142 preguntas, con una media de 15 preguntas por categoría. Se decidió crear un cuestionario de alternativa múltiple con cuatro posibilidades de respuesta, de acuerdo con los estudios que consideran que los distractores son funcionales si hay entre tres y cinco opciones (Downing, 2006; Haladyna, 2004; Haladyna y Downing, 1993). Así pues, siguiendo la línea de estos estudios, se dispuso, para todas las preguntas, una respuesta correcta, otra claramente incorrecta y dos incorrectas, pero que pretendían inducir a error. Para elaborar las preguntas, se escogieron actividades de libros de texto de cuatro colecciones de amplia tirada nacional y de la literatura especializada en didáctica de la lengua (Prado, 2004; Mendoza, 2003).

Las investigadoras seleccionaron 40 preguntas de ese banco inicial en atención al criterio de representatividad y presencia en el currículum. Así, la distribución de preguntas por categoría en el cuestionario resultante fue la siguiente: morfología (ítems 1, 2, 3, 4); fonética, fonología y ortografía (ítems 5, 6, 7, 8); sintaxis (ítems 9, 10, 11, 12); variedad lingüística y sociocultural (ítems 13, 14), nivel léxico-semántico (ítems 15, 16, 17, 18), recursos literarios (ítems 19, 20, 22, 23); tipología textual (ítems 24, 25, 26, 27); procesos de expresión oral y escrita/

pragmática (ítems 28, 29, 30, 31, 32, 33, 34, 35, 36, 37); procesos de comprensión oral y escrita (ítems 38, 39, 40, 41). Se incluyó, además, una pregunta de control de la atención (ítem 21). En el apéndice, se pueden consultar ejemplos de preguntas y respuestas posibles para cada categoría. La puntuación para cada ítem fue 0 (opción incorrecta) y 1 (opción correcta); las puntuaciones para cada ítem y para el total del cuestionario se obtuvieron calculando el valor medio de los ítems implicados.

En la fase 3, este cuestionario se pasó a una muestra de 53 estudiantes de segundo curso. Se les preguntó sobre la comprensibilidad y el grado de dificultad de las preguntas y respuestas, así como sobre la pregunta de control. Con los datos de esta muestra, se analizó la capacidad discriminativa de los ítems y su complejidad (Hurtado, 2018).

En la fase 4, con el cuestionario modificado de 40 preguntas, se hizo un primer pase con la muestra del estudio para evaluar la validez de constructo y un segundo pase para valorar la fiabilidad. En ambos casos, se emplearon las puntuaciones de los ítems, la media de los ítems de las categorías finales y la media total de los ítems por curso (Tabla 1). Todas las puntuaciones medias calculadas en el estudio se distribuyeron normalmente.

Con los ítems resultantes, se llevó a cabo un estudio de validez de constructo basado en el análisis factorial por pasos según los índices alfa de Cronbach (Taber, 2018); como resultado, se eliminaron y se promediaron aquellos ítems que indicó el

modelo. Para el análisis de validez de constructo, se utilizaron los valores promedio de las variables recogidas en el primer pase agrupadas según los resultados del modelo. Este análisis se aplicó también a las agrupaciones de estudiantes por curso para evaluar su adecuación a dichas variables.

Para el estudio de la fiabilidad, se aplicó un análisis psicométrico, basado en el diseño test-retest, a las variables tomadas en dos momentos diferentes (T1 y T2). Se calculó la diferencia entre las puntuaciones y la desviación estándar de la diferencia; se aplicó el coeficiente de correlación intraclass (CCI) (Shrout y Fleiss, 1979) sobre los valores promedios de cada uno de los temas en los diferentes momentos (T1 y T2) con los intervalos de confianza (95%); también el error estándar de medida, el coeficiente de repetibilidad y el cambio mínimo detectable (Beckerman *et al.*, 2001; Bland y Altman, 1986). Los valores del CCI se evaluaron según las indicaciones de estudios previos (Landis y Koch, 1977).

Para estudiar el error de medida, se utilizó el *plot* de Bland-Altman (Bland y Altman, 1996). Para analizar el promedio del error de la diferencia, se calcularon los límites de acuerdo (95%) y sus intervalos de confianza (Bland y Altman, 2010). Para conocer si los valores de error entre los pases fueron significativos, se aplicó la prueba *t* para una muestra sobre las diferencias de los promedios de T1 y T2.

El estudio de la evolución del error de medida con relación a los valores promedios T1 y T2 se calculó mediante un análisis de regresión (Bland y Altman, 1986). El efecto

suelo/techo de las puntuaciones se calculó comparando los porcentajes de participantes con valores de primer y último cuartil de las puntuaciones obtenidas en el primer pase. En los casos en que más del 15% de la población de estudio se encontraba en alguno de estos cuartiles, se consideró que el efecto suelo o techo se daba en el uso de esta herramienta (Terwee *et al.*, 2007).

Con el fin de descartar la posibilidad de un efecto de la variable sexo en el comportamiento de las puntuaciones de los sujetos del estudio, se aplicó un análisis de varianza de medidas repetidas en las puntuaciones de T1 y T2 con el análisis del factor sexo en el tiempo (T1 - T2).

El estudio de la dificultad del cuestionario y de los ítems se calculó con los porcentajes de puntuación de la muestra sobre el valor total. Este valor también se analizó por cursos.

Para evaluar el conocimiento en lengua de la muestra, se calcularon las medias y desviaciones de todos los sujetos por curso, para cada categoría y para las puntuaciones totales. Para analizar el efecto de los diferentes cursos en las puntuaciones de las categorías y en la puntuación total, se aplicó un ANOVA de un factor (curso). Por último, para los contrastes *post hoc*, se aplicó la prueba de Bonferroni.

3. Resultados

3.1. Validez del contenido

Tras consultar, en la fase 1, la bibliografía especializada, el documento

puente y las categorías establecidas por Muedra (2020), en la fase 2, se confeccionaron un total de 142 preguntas, de las cuales, tras un cribado, se enviaron 40 a los expertos.

Estos determinaron por unanimidad que no faltaba ningún contenido imprescindible y señalaron que el cuestionario evaluaba los conocimientos básicos que debe poseer un maestro. Hicieron sugerencias sobre la redacción de algunos ítems para reducir ciertas ambigüedades o ajustar el nivel de dificultad; en concreto, propusieron incrementar el nivel de dificultad de algunas respuestas incorrectas.

Como resultado, se hicieron modificaciones en 18 ítems. Para asegurar la comprensibilidad y pertinencia de los cambios, se pidió a dos expertos que evaluaran de nuevo el cuestionario. Se modificaron 4 ítems en relación con la redacción y el nivel de dificultad de las respuestas.

En cuanto a los datos procedentes de la prueba piloto (fase 3), los estudiantes explicaron que el cuestionario, en general, les pareció concreto y comprensible. Con respecto a la pregunta de control, no la habían detectado como tal debido a que les pareció que formaba parte del contenido; en consecuencia, fue sustituida.

3.2. Validez del constructo y fiabilidad

En la fase 4, del análisis de la capacidad discriminatoria de los 40 ítems de la muestra de estudio, 24 obtuvieron un índice de discriminación bajo ($<.125$), 14 presentaron un índice de dificultad muy bajo ($>93\%$ aciertos), y 2, un índice

de dificultad muy alto (<10 % aciertos). Los ítems que cumplieron con las dos condiciones (baja capacidad de discriminación y dificultad muy alta/baja) fueron eliminados. Así pues, se suprimieron 10 de ellos y quedó un cuestionario de 30.

El análisis factorial por pasos en el estudio de la escala como un único factor determinó la falta de consistencia de estos 30 ítems. El modelo indicó qué ítems reducían la consistencia interna y debían ser eliminados. El resto de los ítems fueron agrupados en tres categorías: 3 ítems de morfología, 3 del nivel léxico-semántico y 3 de sintaxis formaron la categoría MORF_LEX_SINT; 3 ítems de fonética y fonología y 3 de recursos literarios formaron la categoría FFO_RECLIT; y 3 ítems de

tipología textual, 6 de procesos de expresión oral y escrita y 4 de procesos de comprensión oral y escrita formaron la categoría TT_PROEX_PROCOM. Con los promedios de estos 28 ítems agrupados en tres categorías, el instrumento logró una buena consistencia interna (alfa Cronbach = .74).

El análisis de consistencia en función del sexo fue .71 (hombres) y .76 (mujeres). El estudio por grupos indicó un índice .75 para estudiantes de primero, .75 para los de segundo, .67 para los de cuarto y .65 para los de máster.

Los valores medios de las puntuaciones para ambos momentos fueron ligeramente superiores a la mediana de la escala (Tabla 1). Las puntuaciones mejoraron en T2 de forma generalizada.

TABLA 1. Valores del test-retest para las puntuaciones del cuestionario ($n = 190$).

	M T1 (±SD)	M T2 (±DS)	M T1_T2 (±DS)	Dif. M T2 - T1 (±DS)	R	ICC (CI.95%)	CR	SEM	MDC
Total	.64(.12)	.67(.13)	.66(.12)	.03(.10)**	.69**	.81 (.75-.86)	.20	.04	.12
MORF LEX_SINT	.66(.17)	.68(.18)	.67(.16)	.02(.17)	.52*	.68 (.58-.76)	.34	.10	.27
FF RECLIT	.63(.23)	.67(.21)	.65(.19)	.03(.21)*	.52*	.68 (.58-.76)	.42	.12	.34
TT PROEX PROCOM	.64(.16)	.67(.16)	.65(.14)	.03(.15)*	.53*	.69 (.59-.77)	.30	.09	.24

Nota: M = media, T1 = tiempo 1, T2 = tiempo 2, DS = desviación estándar, R = coeficiente de correlación, ICC = coeficientes de correlación intraclase, CI = intervalo de confianza, CR = coeficiente de repetibilidad, SEM = error estándar de medida, MDC = cambio mínimo detectable; diferencia significativa: * $p < .05$, ** $p < .01$.

El error promedio para las puntuaciones totales fue muy pequeño (.03) y el SEM (.04) mostró un bajo error de medida, con valores ligeramente superiores a las diferencias de las medias e inferiores a la DS de la diferencia. Esto ocurrió de la misma forma en las agrupaciones de los ítems. El CR también tuvo un buen comportamiento y proporcionó valores iguales o inferiores a dos veces el DS de la diferencia. El MDC indicó valores de sensibilidad del instrumento muy ajustados y mostró cambios verdaderos en el uso del instrumento a partir de valores de 0.12 puntos en la puntuación total.

Los fuertes coeficientes de correlación intraclase para las puntuaciones totales del test-retest (Tabla 1) indicaron una excelente fiabilidad de las medidas en el tiempo. Sin embargo, se observaron diferencias significativas entre las medidas de los dos pases en las puntuaciones totales y en dos de las tres agrupaciones de los ítems.

Las figuras 1 y 2 muestran los valores absolutos y relativos de las diferencias de las puntuaciones en función de sus va-

lores medios. El valor de la media de las diferencias fue de .03 (DS .10) (Figura 1), equivalente a un porcentaje de error del 3.68% (Figura 2); es decir, no supera el 5% de probabilidad de error asumible. El análisis de regresión reveló que las diferencias entre el test y el retest no cambiaron en la misma medida que las medias de las puntuaciones de ambos tiempos ($F_{(1,189)} = .2; p = .656; \text{beta} = .03$), lo que indicó que las diferencias encontradas entre las puntuaciones del T1 y T2 no variaron en los diferentes niveles de conocimiento de la muestra estudiada.

El tiempo medio utilizado por la muestra para responder el cuestionario fue de 14.67 minutos (DS 4.06).

No se observó un efecto suelo/techo en las puntuaciones promedio logradas por los participantes en el uso de este cuestionario. Ningún sujeto obtuvo puntuaciones promedio por debajo de .34 ni por encima de .89. Sin embargo, el 23% de los sujetos tuvieron puntuaciones en el último cuartil.

FIGURA 1. *Plot* de Bland Altman en valores absolutos de las puntuaciones.

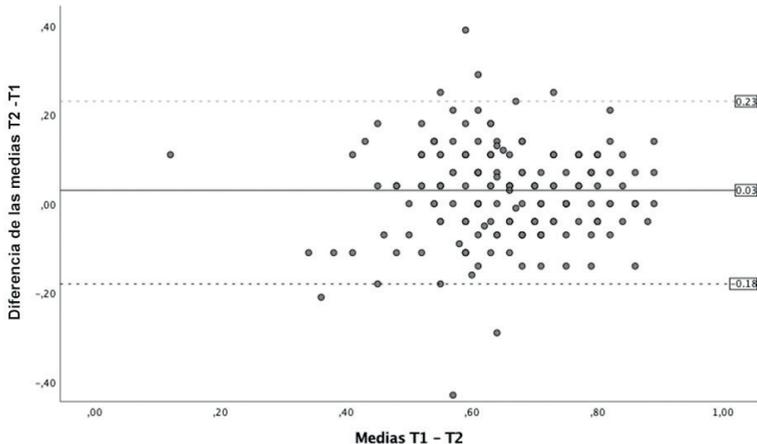
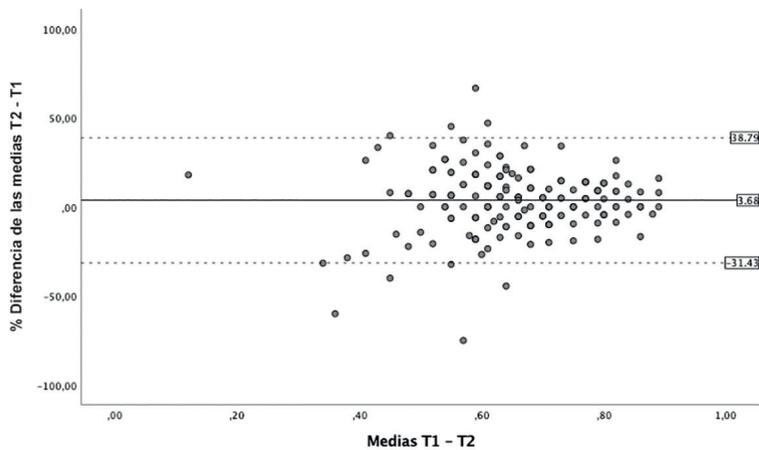


FIGURA 2. Plot de Bland Altman en valores relativos de las puntuaciones.



Este estudio del error de medida también se aplicó a la muestra agrupada por curso y por género. Con relación a los cursos, los estudiantes con error de medida por debajo del 5 % fueron los de cuarto curso (0.8%) y máster (2.04%), mientras que el porcentaje de error en los de primer curso fue del 5.89%, y en los de segundo curso, del 7.54%. Respecto al género, los resultados indicaron un porcentaje de error del 4.58% en hombres y del 5.98% en mujeres.

El ANOVA para medidas repetidas indicó que el sexo no influyó en los cambios que se produjeron con el transcurso del tiempo entre las dos medidas ($F_{(1,126)} = 2.80; p = .598$).

El nivel de dificultad del cuestionario para la muestra de estudio fue medio. El 17.86% no superó la prueba, el 50% obtuvo una puntuación entre 5 y 7, y el 32.14% superó la puntuación de 7. Estos porcentajes variaron entre los diferentes grupos. El mayor porcentaje de estudiantes que no superaron la prueba estuvo en el grupo de

cuarto curso (24.14%). El grupo que obtuvo mayor porcentaje de estudiantes en los niveles más altos de puntuación (>7) fue el de máster (60%).

El análisis de la dificultad de los ítems indicó que los más complejos, con índices de error superiores a los de aciertos, fueron 5 (SIN09, LEX16, RECLIT20, PROCOM38, PROCOM41). Los ítems que tuvieron un índice medio de dificultad, con un porcentaje de aciertos entre el 50% y el 70%, fueron 14 (MORF04, FF05, FF07, SIN10, SIN11, RECLIT22, RECLIT23, TT24, TT25, PROEX29, PROEX30, PROEX31, PROEX34, PROCOMP39). Los ítems más fáciles para la muestra, con porcentajes de acierto de hasta el 80%, fueron 5 (MORF1, MORF3, FF08, PROEX28, PROCOMP40). Por último, se contabilizaron 4 ítems con porcentajes superiores al 80% (hasta el 90%, LEX15 y PROEX32; por encima del 90%, LEX17 y TT26).

3.3. Nivel de conocimiento de la lengua española de los estudiantes en formación

En la Tabla 1, se observa que el nivel de conocimiento promedio fue de 0.64 puntos, lo que indica que la muestra es-

tudiada alcanzó una puntuación media de conocimientos en la puntuación total y en las diferentes categorías. En la Tabla 2, se recogen los resultados de la muestra segregados por curso.

TABLA 2. Puntuaciones promedio total y por categorías segregadas por curso.

Categoría	Curso	N	Media	DS
MORF_LEX_SINT_01	Primero	65	0.658	0.169
	Segundo	40	0.671	0.180
	Cuarto	65	0.634	0.165
	Máster	20	0.743	0.191
FF_RECLIT_01	Primero	65	0.633	0.213
	Segundo	40	0.625	0.238
	Cuarto	65	0.595	0.228
	Máster	20	0.762	0.199
TT_PROEX_PROCOM_01	Primero	65	0.615	0.169
	Segundo	40	0.652	0.154
	Cuarto	65	0.645	0.152
	Máster	20	0.694	0.134
Total_01	Primero	65	0.632	0.128
	Segundo	40	0.652	0.132
	Cuarto	65	0.630	0.114
	Máster	20	0.724	0.104

Nota: DS = desviación estándar.

Del estudio de las frecuencias de respuesta correctas e incorrectas, se infiere que el 12.6% suspendió la prueba (<5), el 49.8% obtuvo entre el 5 y el 7, y el 37.2% alcanzó una buena calificación (>7).

En cuanto al análisis del efecto de los diferentes cursos, los resultados indican que las puntuaciones fueron diferentes solo para la variable cursos en la media to-

tal ($F_{(3,189)} = 3,408; p = .019$; eta cuadrado = .045) y para la categoría FF_RECLIT ($F_{(3,189)} = 2,902; p = .036$; eta cuadrado = .052). El análisis *post hoc* señala que las diferencias se produjeron entre los cursos de primero y máster (dif. -.092; $p = .022$) y entre los de cuarto y máster (dif. -.094; $p = .017$) para la puntuación total, y entre los cursos de cuarto y máster (dif. -.17; $p = .022$) para la categoría FF_RECLIT.

4. Discusión

El objetivo de este trabajo era diseñar y validar un instrumento que permitiera evaluar el CC en lengua española de los maestros en formación. En línea con las investigaciones en otras disciplinas (Verdugo *et al.*, 2019; Vásquez y Alsina, 2015; Godino *et al.*, 2016), se partió de la literatura especializada, del marco normativo y de los libros de texto para generar un banco de preguntas que fue sometido a un proceso de validación de contenido y de constructo.

Se recurrió al método Delphi a fin de garantizar niveles de validez altos; para evitar sus posibles desventajas, se siguió de forma escrupulosa tanto sus características, el desarrollo de sus fases y la selección de expertos (Cabero e Infante, 2014) como su uso en la población diana. El método Delphi resulta especialmente útil a la hora de diseñar y validar un instrumento cuando no existe ninguno que se ajuste a las necesidades de la investigación (Andrés *et al.*, 2019). Si bien existen algunas herramientas que miden el conocimiento que deben tener los maestros para enseñar a leer en el ámbito anglosajón (Wahsburn *et al.*, 2016; Binks-Cantrell *et al.*, 2012), el presente estudio aporta el primer instrumento que permite evaluar el CC en lengua española de los futuros maestros. El cuestionario resultante, compuesto de 28 preguntas más la de control, se muestra válido y fiable para este propósito.

En cuanto al contenido, expertos y usuarios fueron unánimes a la hora de afirmar su pertinencia; con respecto a la formulación de preguntas y respuestas, se llevaron a cabo las modificaciones que se

consideraron pertinentes de las sugeridas por ambos grupos. El pase al grupo piloto sirvió también para llevar a cabo un primer análisis discriminatorio.

En relación con la validez de constructo, tras el proceso de eliminación y agrupación de ítems, el instrumento alcanzó una buena consistencia interna tanto global como para los diferentes sexos y cursos, si bien su consistencia fue menor en los cursos más altos. Esto indica un funcionamiento del cuestionario robusto con independencia del género y del curso.

El análisis de la fiabilidad de las medidas estuvo respaldado por los valores medios de las puntuaciones en el test-retest, la fuerte asociación entre las medidas en los diferentes momentos y el estrecho margen del error de medida. Aunque se observa que los valores promedio mejoraron de forma significativa en el retest en las agrupaciones FF_RECLIT y TT_PROEX_PROCOM (grupos en los que se mostraron más inestables), el análisis de regresión mostró que las diferencias encontradas entre las puntuaciones no variaron en la medida en que los valores promedio aumentaron. En consecuencia, se confirma que no se produjo un efecto distorsionador entre medidas debido al aprendizaje en el proceso.

Desde el punto de vista metodológico, estos resultados corroboran las afirmaciones de Bland y Altman (1996): los estudios de fiabilidad con un análisis de correlación de los ítems en los diferentes momentos proporcionan información insuficiente sobre su estabilidad en el tiempo. Tales estudios de fiabilidad requieren profundizar en

el análisis del error de medida (y, en este sentido, examinar la fiabilidad relativa y la absoluta) para confirmar el efecto del tiempo en el uso del instrumento (Vaz *et al.*, 2013). Por una parte, la precisión de los índices que este ofrece a la hora de medir el conocimiento no es equiparable a la de estudios de variables más objetivas; esto es habitual cuando se utilizan herramientas que evalúan conceptos complejos como los que se abordan en este trabajo. Sin embargo, como indica el análisis de regresión, sí mantiene su nivel de precisión con independencia del grado de conocimiento de los estudiantes. A este resultado sobre la fiabilidad de la herramienta hay que añadir que el promedio del error de medida es muy bajo (0.03), lo que equivale a un error asumible, es decir, a un porcentaje de probabilidad de medidas diferentes no significativo (3.68%). La probabilidad de que ocurra una diferencia significativa entre las medidas no alcanza el 5%.

A los buenos resultados del error de medida hay que sumar un comportamiento psicométrico de la escala óptimo, en especial en los valores totales (Tabla 1). Con independencia de las agrupaciones de los ítems en tres categorías, la escala ha sido evaluada como una escala unifactorial que se refiere de forma general al conocimiento en lengua castellana del profesorado en formación. La variabilidad de la medida del instrumento (SEM = 0.04) fue similar al error de medida relativo (0.03). El error absoluto de medida o CR nos indicó que las variaciones en la medida superiores a 0.20 puntos serían medidas cuyo valor superaría el error absoluto teórico del instrumento y se podrían

considerar variaciones verdaderas. También que los cambios en la puntuación del cuestionario iguales o superiores al valor de MDC (0.12) se podrían considerar modificaciones reales en el conocimiento de los estudiantes.

Aunque la fiabilidad del cuestionario es buena, los datos que se extraigan de su uso deben interpretarse con cautela, ya que el instrumento se mostró menos fiable en el tiempo en el análisis del subgrupo de mujeres y en los grupos con menor experiencia (primero y segundo). En este sentido, el estudio ha profundizado en el análisis del comportamiento del error de medida en función del sexo y de los diferentes niveles de formación de los estudiantes; así, el test funcionó mejor en los grupos de cuarto y de máster; así como en el de género masculino. Estas agrupaciones tuvieron situaciones especiales que podrían justificar tales resultados. Por un lado, el 80% de la muestra fueron mujeres, lo que podría explicar la mayor dispersión de las puntuaciones entre los registros del test y del retest, y, en consecuencia, el mayor error de medida. Por otro, las diferencias entre los grupos de mayor y menor experiencia pudieron deberse al hecho de que los estudiantes de cursos más altos tuviesen un nivel de conocimientos más consolidado, con independencia de si este era mayor o menor.

En el estudio de usabilidad, el cuestionario se mostró manejable y útil para la formación del profesorado: el tiempo medio para responder de 15 minutos, aproximadamente, y no resultó difícil de comprender.

En cuanto a los resultados sobre el nivel de conocimientos, se observa que el 12.6% suspendió (<5), el 49.8% obtuvo entre el 5 y el 7, y el 37.2% alcanzó una buena calificación (>7). Desde una perspectiva matemática, podría considerarse que la distribución de las puntuaciones se encuentra en un nivel de conocimientos medio aceptable (6-7 puntos), con una puntuación media de .64 y con el 87% de estudiantes que superaron el test. No obstante, cabe recordar que este se diseñó con preguntas de conocimientos básicos, procedentes en su mayoría de los libros de primaria. Por tanto, no era esperable que casi el 50% de la muestra se encontrase entre las puntuaciones 5 y 7, si bien es cierto que carencias similares se han puesto de manifiesto en los estudios de otras áreas (Verdugo *et al.*, 2019; Vásquez y Alsina, 2015; Depaepe *et al.*, 2013). Lo que resalta este estudio es que la muestra tiene un claro déficit en contenido disciplinar en lengua española.

Es preocupante, además, que casi el 13% de estos futuros profesionales no supere la prueba. El CC del profesorado está altamente relacionado con el aprendizaje de los estudiantes, por lo que cumplir con el requisito de conocer un contenido para poder enseñarlo (Friedrichsen *et al.*, 2009) es una responsabilidad tanto de los formadores de los estudiantes de Magisterio como de las instituciones públicas implicadas.

Por otra parte, no es esperable que los estudiantes de cuarto curso sean los que menos puntuación obtengan. Si bien la tendencia general del conocimiento aumenta a medida que se avanza en el grado, los niveles de significancia indican que la

evolución del conocimiento no es significativa. Ello podría deberse a que las materias disciplinares se cursan sobre todo en los dos primeros cursos y son sustituidas en los dos últimos por las materias específicamente didácticas. Las diferencias con los estudiantes de máster son esperables, dado que estos tienen una formación disciplinar específica más amplia.

En cuanto a las limitaciones de este trabajo, debemos exponer que no ha podido aplicarse un análisis de la validez de convergencia o de criterio, dado que no existen instrumentos equiparables. La muestra de estudio en el proceso de validación la consideramos adecuada; sin embargo, se deberían utilizar otras muestras con características culturales diferentes.

5. Conclusiones

Tras el análisis de contenido y fiabilidad, cabe afirmar que el instrumento presentado es válido y fiable para medir el CC en lengua española de los maestros en formación.

En apariencia, los primeros datos indican que los estudiantes tienen un conocimiento aceptable de lengua. No obstante, si se tiene en cuenta que el cuestionario pretende medir conocimientos mínimos, resulta llamativo que la mitad de la muestra no obtengan más que lo que equivaldría a un aprobado alto o notable.

La siguiente fase de esta investigación, que consistirá en la aplicación de este instrumento a grandes muestras de población, aclarará si este hecho se debe

simplemente a las dimensiones de la muestra o, por el contrario, desvela una realidad preocupante sobre la formación de los maestros, hipótesis que parece estar refrendada por la investigación en otras áreas. Instrumentos como este pueden contribuir a definir y subsanar estas posibles deficiencias, mediante la aportación de datos consistentes a los formadores de maestros para una orientación más efectiva de sus prácticas.

Apéndice. Ejemplos de preguntas del cuestionario definitivo

La respuesta correcta aparece marcada en cursiva.

MORF_03. Indica cuál de estas oraciones NO contiene ningún verbo en modo subjuntivo:

- a) Quizá Teresa y Silvia lleguen tarde al partido.
- b) Ojalá que llueva más esta primavera.
- c) Si tuvieras más interés, estudiar te resultaría más fácil.
- d) *Felipe participará en la carrera el domingo con su padre.*

FFO_05. Desde el punto de vista ortográfico, ¿cuál de estas oraciones es correcta?

- a) *Dime qué te pasa hoy*
- b) No sé donde vive Paquita.

- c) He olvidado cuando tengo cita con el médico.
- d) No sé en que momento dejó de dolerme.

LEX_17. Elige la opción en que todas las palabras sean derivadas:

- a) *Imperial, combativo, volcánico, montañoso.*
- b) Combativo, volcánico, amor, limón.
- c) Volcánico, montañoso, mesa, corazón.
- d) Amor, limón, mesa, corazón.

RECLIT_20. Elige la afirmación verdadera:

- a) Un soneto tiene rima asonante.
- b) *Un soneto tiene 14 versos.*
- c) Un soneto puede ser de arte mayor o menor.
- d) Un soneto puede tener un número ilimitado de estrofas.

TT_26. Señala la opción que contenga únicamente géneros orales:

- a) *El diálogo, el debate, la rueda de prensa y el coloquio.*
- b) La entrevista, la exposición, el recetario y la noticia.
- c) El diario personal, la biografía, el libro de viajes y la descripción.

- d) El diálogo, el debate, el diario personal y el coloquio.

PROEX_31. Cuál de estas afirmaciones NO corresponde a la planificación de la escritura:

- a) Hacer una lluvia de ideas.
 b) *Corregir la ortografía.*
 c) Búsqueda de modelos.
 d) Hacer un esquema.

Referencias bibliográficas

Alfaro-Carvajal, C., Flores-Martínez, P., y Valverde-Soto, G. (2022). Conocimiento de profesores de matemáticas en formación inicial sobre la demostración: Aspectos lógico-matemáticos en la evaluación de argumentos. *Uniciencia*, 36 (1), 140-165. <https://doi.org/10.15359/ru.36-1.9>

Andrés, I., Muñoz, M., Ruíz, G., Gil, B., Andrés, M., y Almaraz, A. (2019). Validación de un cuestionario sobre actitudes y práctica de actividad física y otros hábitos saludables mediante el método Delphi. *Revista Española de Salud Pública*, 93.

Beckerman, H., Roebroek, M. E., Lankhorst, G. J., Becher, J. G., Bezemer, P. D., y Verbeek, A. (2001). Smallest real difference, a link between reproducibility and responsiveness [La diferencia real más pequeña, un vínculo entre la reproducibilidad y la capacidad de respuesta]. *Quality of Life Research*, 10 (7), 571-578. <https://doi.org/10.1023/A:1013138911638>

Binks-Cantrell, E., Joshi, R. M., y Washburn, E. K. (2012). Validation of an instrument for assessing teacher knowledge of basic language constructs of literacy [Validación de un instrumento para evaluar el conocimiento del profesorado sobre constructos lingüísticos básicos de alfabetización]. *Annals of dyslexia*, 62, 153-171. <https://doi.org/10.1007/s11881-012-0070-8>

Bland, J. M., y Altman, D. (1986). Statistical methods for assessing agreement between two methods of

clinical measurement [Métodos estadísticos para evaluar la concordancia entre dos métodos de medición clínica]. *The Lancet*, 327 (8476), 307-310. [https://doi.org/10.1016/S0140-6736\(86\)90837-8](https://doi.org/10.1016/S0140-6736(86)90837-8)

Bland, J. M., y Altman, D. G. (1996). Statistics notes: Measurement error [Notas estadísticas: error de medida]. *Bmj*, 312 (7047), 41-42. <https://doi.org/10.1136/bmj.312.7047.1654>

Bland, J. M., y Altman, D. G. (2010). Statistical methods for assessing agreement between two methods of clinical measurement [Métodos estadísticos para evaluar la concordancia entre dos métodos de medición clínica]. *International journal of nursing studies*, 47 (8), 931-936. <https://doi.org/10.1016/j.ijnurstu.2009.10.001>

Borg, S. (2003). Teacher cognition in grammar teaching: A literature review [La cognición del profesor en la enseñanza de la gramática: una revisión bibliográfica]. *Language awareness*, 12 (2), 96-108. <https://doi.org/10.1080/09658410308667069>

Bucat, R. (2005). Implications of chemistry education research for teaching practice: Pedagogical content knowledge as a way forward [Implicaciones de la investigación en educación química para la práctica docente: el conocimiento pedagógico del contenido como vía de progreso]. *Chemistry Education International*, 6 (1), 1-2.

Cabero, J., e Infante, A. (2014). Empleo del método Delphi y su empleo en la investigación en Comunicación y Educación. *EDUTEC Revista Electrónica de Investigación Educativa*, (48), 1-16. <https://doi.org/10.21556/edutec.2014.48.187>

CEFIRE. (2015). *Documento puente. Lengua española. Comunitat Valenciana*. https://drive.google.com/file/d/1UnWPGNgG_v7-UnzX42BP746IHFC-HytD/view?usp=sharing

Chetty, R., Friedman, J. N., y Rockoff, J. E. (2011). *The long-term impacts of teachers: Teacher value-added and student outcomes in adulthood [El impacto a largo plazo de los profesores: el valor añadido del profesor y los resultados de los alumnos en la edad adulta]*. National Bureau of Economic Research. <https://doi.org/10.3386/w17699>

Depaepe, F., Verschaffel, L., y Kelchtermans, G. (2013). Pedagogical content knowledge: A systematic review of the way in which the concept has pervaded mathematics educational research [Conocimiento pedagógico del

- contenido: una revisión sistemática del modo en que el concepto se ha extendido en la investigación educativa en matemáticas]. *Teaching and teacher education*, 34, 12-25. <https://doi.org/10.1016/j.tate.2013.03.001>
- Downing, S. M. (2006). Selected-response item formats in test development [Formatos de ítems de respuesta seleccionada en el desarrollo de tests]. En S. M. Downing, y T. M. Haladyna (Eds.), *Handbook of test development [Manual de desarrollo de pruebas]* (pp. 287-302). Taylor & Francis. <https://doi.org/10.4324/9780203874776>
- Friedrichsen, P. J., Abell, S. K., Pareja, E. M., Brown, P. L., Lankford, D. M., y Volkmann, M. J. (2009). Does teaching experience matter? Examining biology teachers' prior knowledge for teaching in an alternative certification program [¿Importa la experiencia docente? Análisis de los conocimientos previos de los profesores de biología en un programa de certificación alternativo]. *Journal of Research in Science Teaching*, 46 (4), 357-383. <https://doi.org/10.1002/tea.20283>
- Gess-Newsome, J., Taylor, J. A., Carlson, J., Gardner, A. L., Wilson, C. D., y Stuhlsatz, M. A. (2019). Teacher pedagogical content knowledge, practice, and student achievement [Conocimiento pedagógico del contenido, práctica y rendimiento de los alumnos]. *International Journal of Science Education*, 41 (7), 944-963. <https://doi.org/10.1080/09500693.2016.1265158>
- Godino, J. D., Gonzato, M., Contreras, Á., Estepa, A., y Díaz-Batanero, C. (2016). Evaluación de conocimientos didáctico-matemáticos sobre visualización de objetos tridimensionales en futuros profesores de educación primaria. *Journal of Research in Mathematics Education*, 5 (3), 235-262. <https://doi.org/10.17583/redimat.2016.1984>
- Haladyna. (2004). *Developing and validating multiple-choice test items* [Desarrollo y validación de ítems de tests de opción múltiple]. Routledge. <https://doi.org/10.4324/9780203825945>
- Haladyna, T. M., y Downing, S. M. (1993). How many options is enough for a multiple-choice test item? [¿Cuántas opciones son suficientes para una prueba tipo test?]. *Educational and Psychological Measurement*, 53 (4), 999-1010. <https://doi.org/10.1177/0013164493053004013>
- Harris, D. N., y Sass, T. R. (2011). Teacher training, teacher quality and student achievement [Formación del profesorado, calidad del profesorado y rendimiento de los alumnos]. *Journal of public economics*, 95 (7-8), 798-812. <https://doi.org/10.1016/j.jpubeco.2010.11.009>
- Hurtado, L. L. (2018). Relación entre los índices de dificultad y discriminación. *Revista digital de investigación en docencia universitaria*, 12 (1), 273-300. <http://dx.doi.org/10.19083/ridu.12.614>
- Kleickmann, T., Richter, D., Kunter, M., Elsner, J., Besser, M., Krauss, S., y Baumert, J. (2013). Teachers' content knowledge and pedagogical content knowledge: The role of structural differences in teacher education [Conocimiento del contenido de los profesores y conocimiento pedagógico del contenido: el papel de las diferencias estructurales en la formación del profesorado]. *Journal of teacher education*, 64 (1), 90-106. <https://doi.org/10.1177/0022487112460398>
- Landis, J. R., y Koch, G. G. (1977). The measurement of observer agreement for categorical data [Medición del acuerdo entre jueces para datos categóricos]. *Biometrics*, 33, (1), 159-174. <https://doi.org/10.2307/2529310>
- Mendoza, A. (2003). *Didáctica de la lengua y la literatura para educación primaria*. Pearson Educación.
- Ministerio de Educación, Cultura y Deporte. (2012). *TEDS-M. Informe español. Estudio internacional sobre la formación inicial en matemáticas de los maestros*. <https://www.educacionyfp.gob.es/dctm/inee/internacional/tedsmlinea.pdf?documentId=0901e72b8143866e>
- Mokkink, L. B., Terwee, C. B., Patrick, D. L., Alonso, J., Stratford, P. W., Knol, D. L., y De Vet, H. C. (2010). The COSMIN checklist for assessing the methodological quality of studies on measurement properties of health status measurement instruments: an international Delphi study [La lista de comprobación COSMIN para evaluar la calidad metodológica de los estudios sobre las propiedades de los instrumentos de medida del estado de salud: un estudio Delphi internacional]. *Quality of life research*, 19, 539-549. <https://doi.org/10.1007/s11136-010-9606-8>
- Muedra, M.ª T. (2020). *Análisis de la competencia oral en los libros de texto de lengua española de educación primaria* [Trabajo Final de Máster]. Universitat de València.
- Opfer, V. D., y Pedder, D. (2011). Conceptualizing teacher professional learning [Conceptualizar

- el aprendizaje profesional de los profesores]. *Review of educational research*, 81 (3), 376-407. <https://doi.org/10.3102/0034654311413609>
- Pajares, M. F. (1992). Teachers' beliefs and educational research: Cleaning up a messy construct [Creencias de los profesores e investigación educativa: clarificación de un constructo confuso]. *Review of educational research*, 62 (3), 307-332. <https://doi.org/10.3102/00346543062003307>
- Prado, J. (2004). *Didáctica de la lengua y la literatura para educar en el siglo XXI*. La Muralla.
- Shrout, P. E., y Fleiss, J. L. (1979). Intraclass correlations: Uses in assessing rater reliability [Correlaciones intraclase: usos en la evaluación de la fiabilidad de los evaluadores]. *Psychological Bulletin*, 86 (2), 420. <https://doi.org/10.1037/0033-2909.86.2.420>
- Shulman, L. (1987). Knowledge and teaching: Foundations of the new reform [Conocimiento y enseñanza: fundamentos de la nueva reforma]. *Harvard educational review*, 57 (1), 1-23. <https://doi.org/10.17763/haer.57.1.j463w79r56455411>
- Taber, K. S. (2018). The use of Cronbach's alpha when developing and reporting research instruments in science education [El uso del alfa de Cronbach en el desarrollo y presentación de informes sobre instrumentos de investigación en la enseñanza de las ciencias]. *Research in science education*, 48(6), 1273-1296. <https://doi.org/10.1007/s11165-016-9602-2>
- Tattoo, M. T., Peck, R., Schwille, J., Bankov, K., Senk, S. L., Rodriguez, M., y Rowley, G. (2012). *Policy, practice, and readiness to teach primary and secondary mathematics in 17 countries: Findings from the IEA Teacher education and development study in mathematics (TEDS-M) [Política, práctica y preparación para enseñar matemáticas en primaria y secundaria en 17 países: resultados del estudio de la IEA sobre formación y desarrollo del profesorado de matemáticas (TEDS-M)]*. IEA. <https://www.iea.nl/publications/study-reports/international-reports-iea-studies/policy-practice-and-readiness-teach>
- Terwee, C. B., Bot, S. D., De Boer, M. R., Van der Windt, Daniëlle A.W.M., Knol, D. L., Dekker, J., y De Vet, H. C. (2007). Quality criteria were proposed for measurement properties of health status questionnaires [Criterios de calidad para las propiedades de medida de los cuestionarios sobre el estado de salud]. *Journal of Clinical Epidemiology*, 60 (1), 34-42. <https://doi.org/10.1016/j.jclinepi.2006.03.012>
- Vásquez, C., y Alsina, A. (2015). Conocimiento didáctico-matemático del profesorado de educación primaria sobre probabilidad: diseño, construcción y validación de un instrumento de evaluación. *Bolema: Boletim de Educação Matemática*, 29 (52), 681-703. <https://doi.org/10.1590/1980-4415v29n52a13>
- Vaz, S., Falkmer, T., Passmore, A. E., Parsons, R., y Andreou, P. (2013). The case for using the repeatability coefficient when calculating test-retest reliability [Razones para utilizar el coeficiente de repetibilidad en el cálculo de la fiabilidad test-retest]. *PLoS One*, 8 (9), e73990. <https://doi.org/10.1371/journal.pone.0073990>
- Verdugo, J. J., Solaz, J. J., y Sanjosé, V. (2019). Evaluación del conocimiento científico en maestros en formación inicial: el caso de la Comunidad Valenciana. *Revista de Educación*, (383), 133-162. <https://doi.org/10.4438/1988-592X-RE-2019-383-404>
- Washburn, E. K., Binks-Cantrell, E. S., Joshi, R. M., Martin-Chang, S., y Arrow, A. (2016). Pre-service teacher knowledge of basic language constructs in Canada, England, New Zealand, and the USA [Conocimientos de los profesores en formación sobre constructos lingüísticos básicos en Canadá, Inglaterra, Nueva Zelanda y EE. UU.]. *Annals of dyslexia*, 66, 7-26. <https://doi.org/10.1007/s11881-015-0115-x>

Biografía de los autores

María Teresa Muedra-Pedris. Estudiante del Programa de Doctorado en Didácticas específicas de la Universidad de Valencia. Es miembro, junto con los demás autores, del grupo de investigación «Conocimiento didáctico del contenido: fundamentos del análisis y la acción didáctica del profesorado (CDC-GIUV2016-280)».

 <https://orcid.org/0000-0002-0592-4857>

Manuel Monfort-Pañego. Profesor en el Departamento de Didáctica de la Educación Física, Artística y Música de la Universidad de Valencia. Doctor en Educación Física por la Universidad de Valencia. Sus líneas de investigación se centran en la formación de maestros, principalmente en el desarrollo y la validación de instrumentos de medida para evaluar conocimientos y hábitos del alumnado en relación con la educación física. Es miembro, junto con los demás autores, del grupo de investigación «Conocimiento didáctico del contenido: fundamentos del análisis y la acción didáctica del profesorado (CDC-GIUV2016-280)».

 <https://orcid.org/0000-0002-3181-2170>

Ángela Gómez-López. Profesora en el Departamento de Didáctica de la Lengua y la Literatura de la Universidad de Valencia. Doctora en Didácticas Específicas por la Universidad de Valencia. Sus líneas de investigación se centran en la formación de maestros: aprendizaje y enseñanza de lenguas extranjeras, comprensión lectora y

control de la comprensión (metacognición) en L2 y análisis didáctico del contenido escolar en lenguas maternas y extranjeras. Es miembro, junto con los demás autores, del grupo de investigación «Conocimiento didáctico del contenido: fundamentos del análisis y la acción didáctica del profesorado (CDC-GIUV2016-280)».

 <https://orcid.org/0000-0001-7527-5007>

Eva Morón-Olivares. Profesora en el Departamento de Didáctica de la Lengua y la Literatura de la Universidad de Valencia. Doctora en Filología Hispánica por la Universidad de Granada. Sus líneas de investigación se centran en la formación de maestros, en especial en la educación literaria y el análisis didáctico del contenido escolar en lenguas maternas y extranjeras. Es miembro, junto con los demás autores, del grupo de investigación «Conocimiento didáctico del contenido: fundamentos del análisis y la acción didáctica del profesorado (CDC-GIUV2016-280)».

 <https://orcid.org/0000-0003-2180-2857>