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

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

María-Teresa MUEDRA-PERIS. Doctoral Student. Universidad de Valencia (*muepe@alumni.uv.es*).
Manuel MONFORT-PAÑEGO, PhD. Associate Professor. Universidad de Valencia (*manuel.monfort@uv.es*).
Ángela GÓMEZ-LÓPEZ, PhD. Associate Professor. Universidad de Valencia (*angela.gomez@uv.es*).
Eva MORÓN-OLIVARES, PhD. Associate Professor. Universidad de Valencia (*eva.moron@uv.es*).

Abstract:

In recent decades, various research works have focussed on assessing teachers' pedagogic knowledge, especially in science and mathematics. However, few works in the field of language have designed and validated tools for this purpose. The present work aims to progress in this line of research by designing and validating a questionnaire to assess the Spanish-language disciplinary knowledge of pre-service teachers. Spanish students from the degree in Primary School Education and Master's in Secondary Education as well as experts in didactics of language and in validating questionnaires participated. To analyse its content validity, we used the Delphi method and, to study consistency, we performed a psychometric analysis using the test–retest reliability method. The instrument was found to be consistent and valid. The results were below what was expected and revealed that the sample showed a clear shortcoming in disciplinary content in Spanish language. These data seem to be in line with those obtained in other areas. Consequently, CoDiLe can contribute to defining and remedying these possible deficiencies by providing consistent data to teacher trainers to guide their practice more effectively.

Keywords: measurement instrument, level of knowledge, Spanish language, teacher training.

Date of receipt of the original: 2023-07-20.
Date of approval: 2023-10-25.
Please cite this article as follows: Muedra-Peris, M. T., Monfort-Pañeco, M., Gómez-López, Á., & Morón-Olivares, E. (2024). CoDiLe: An instrument for evaluating the Spanish-language disciplinary knowledge of pre-service teachers [CoDiLe: un instrumento para evaluar el conocimiento disciplinar de lengua española de los maestros en formación]. *Revista Española de Pedagogía, 82* (288), 271-290. https://doi.org/10.22550/2174-0909.4037

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.

1. Introduction

Assessing a country's educational needs involves considering all of the factors that make up its education system, such as teacher training. This is a complex research topic and is one of the key areas for action to improve the education system; nonetheless, there is no consensus on what factors promote teaching quality and how to incorporate them into initial training (Harris & Sass, 2011): some studies have focussed on professional learning (Opfer & Pedder, 2011), cognition (Borg, 2003), or personal knowledge (Pajares, 1992).

One of the authors who has made the biggest contributions to research on teacher training is Shulman (1987), who proposed a new integrating concept, the *pedagogical content knowledge* (PCK): the combination of content and pedagogy in the comprehension of how certain topics are organised, represented, and adapted to the interests and skills of the students, and how they are presented when they are taught.

This new concept has been a catalyst for significant research works that in recent decades have revealed the difference between *content knowledge* (CK) and its teaching (PCK) (Bucat, 2005). Although the relationship between CK and PCK is not clearly defined in the literature, it does seem to be clear that CK is at the centre of the development of teachers' professional competences (Kleichkman et al., 2013).

In Shulman's model, CK is the first aspect that must be taken into account to study the teaching of disciplines. Research shows that an in-depth CK improves explanations, favours use of resources, and influences students' comprehension and their academic success (Chetty et al., 2011); therefore, it is essential to define what level of knowledge future teachers have in order to act to improve their initial and ongoing training (Kleickmann et al., 2013). Knowing the content of the subject you are going to deliver is a prerequisite for being able to teach it (Friedrichsen et al., 2009). In fact, in some pieces of research, teachers' CK was significant in explaining improvements in students' results (Gess-Newsome et al., 2019).

In recent decades, various research works have focussed on evaluating teachers' CK, especially in the area of science and mathematics. Godino et al. (2016) assessed teachers' knowledge of visualisation of three-dimensional objects in 241 primary teaching students. To design their questionnaire, they used previous research works, the curriculum, and textbooks that are widely used nationally. Although the questions were taken from primary school books, the results showed that 62% of the students did not answer the proposed tasks optimally.

Spain's Ministry of Education, Culture, and Sport (2012) participated in an international study on initial training in mathematics of primary school teachers that used a questionnaire based on the TEDS-M study by Tattoo et al. (2012), which evaluated both their didactic and disciplinary knowledge. To design the questions, previous research and the legal frameworks of the participating countries were used. The final questionnaire had 74 questions in multiple-choice or open-answer format. The mean scores obtained by future teachers from Spain were below the international mean for both mathematical and didactic knowledge, although they did score slightly higher in the latter type.

Vásquez and Alsina (2015) validated a questionnaire with open-ended questions to assess didactic-mathematical knowledge for teaching probability. As the knowledge base, they used previous research, curricular guidelines, and textbooks. Both the pilot application of the instrument and its replication obtained medium-low scores in all categories.

Verdugo et al. (2019) analysed the didactic-disciplinary knowledge of science of pre-service teachers; to do so, they created a questionnaire with 30 multiple-choice items based on Spain's national curriculum and Spanish textbooks. The instrument displayed a command of scientific content with room for improvement and the presence of some significant conceptual errors.

As for CK in language, few works have designed and validated tools to evaluate this; most of them focus on the knowledge needed for teaching how to read. Binks-Cantrell et al. (2012) and Washburn et al. (2016) validated an instrument for evaluating teachers' knowledge of the basic constructs of language that are involved in teaching reading. A total of 279 pre-service teachers participated. The questionnaire included 38 items aimed at content knowledge. The results displayed a lack of knowledge, in particular of morphology and phonology.

The present work seeks to make progress on this line of research: its aim is to design and validate an instrument that makes it possible to evaluate pre-service teachers' content knowledge in Spanish language. As far as we are aware, there are no instruments that enable us to evaluate the disciplinary knowledge that teachers, specifically pre-service ones, must have to deliver Spanish language. The need for this instrument is relevant for research in the field of education for two reasons: because language is disciplinary content and because it is, at the same time, a vehicle for the other types of learning. This need justifies the aim of this work, as creating instruments that make it possible to evaluate the teachers' knowledge of the subjects on the curriculum has a direct effect on teacher training programmes.

2. Methodology

2.1. Study design

To analyse Spanish-language CK of pre-service teachers, we designed and validated a questionnaire using the Delphi method (Andrés et al., 2019) in four stages:

- Stage 1. Evidence collection. Literature search. Selecting evidence indicators.

- Stage 2. Development of version I. Drawing up items. Evaluation by experts.

- Stage 3. Development of version II. Pilot test. Evaluation by students.

- Stage 4. Development of the final version. First pass (construct validity). Second pass (reliability).

To study the reliability of the questionnaire, we used a test-retest process and psychometric analysis.

2.2. Participants

During steps 2 and 3 (content validity), two groups of participants were used to evaluate the content and comprehensibility of the initial test. The first group comprised six independent experts: three from philology and didactics of language, one from PCK, and two from instruments and research designs (one of them also an expert in PCK). The experts were selected in accordance with the following criteria: they had to be outside the study, have a doctorate, be university teachers, and have high-quality publications on didactics of language or PCK, research methods, or validating questionnaires. In parallel, two external subjects who were not related to the content or the selection criteria participated as evaluators.

The second group comprised a natural group of 53 university students (aged 19-23) of both sexes from the second year of the bachelor's degree in Primary Education at a Spanish university. They completed the test online and ten were subsequently selected to be interviewed. They were asked to evaluate the content and comprehensibility of the initial test. The results of the tests were used to make a preliminary estimate of the functioning of the questionnaire.

For stage 4, the sample of participants was 256 students of both sexes (aged 18-25) from the degree programmes in Primary/Early Childhood Education (years 1, 2, and 4) at a Spanish university. A natural group of 20 students (aged 23-35) from the master's in Secondary Education (Spanish Language and its Literature specialism) also participated. This group had more specialised disciplinary knowledge and was used to ascertain whether the instrument discriminated between different levels of knowledge. Of the initial sample, 190 students (152 women and 38 men) completed both passes of the questionnaire.

2.3. Procedure

In stage 1, we used various sources to design the questionnaire. We performed a search for specialist literature on Spanish-language disciplinary knowledge and on the use of questionnaires for evaluating it in the teaching of Spanish as a first and second language.

In stage 2, this analysis was used as a basis for generating a bank of questions and a first version of the questionnaire was agreed on. This was sent through a virtual platform to six independent experts and to two external subjects for its evaluation.

In stage 3, the resulting instrument was tested on a group of 53 students to make a first estimate of its functioning and of the pertinence of the questions. To do so, an online platform was used during a class session. The time was limited to 40 minutes. After this, we interviewed 10 students to complete the information obtained.

Some questions were reformulated or replaced following the suggestions of the two groups. The new version was again sent to two experts: one in questionnaires and research methods, and the other in didactics of Spanish language, and their proposals were also incorporated into the questionnaire.

Finally, during stage 4, the definitive questionnaire was administered to the study sample in two passes: we used the data from the first pass to evaluate the construct validity, and used these data and the data from the second pass to study the reliability. Four experimental conditions were used to counterbalance the order of the questions in the first and second pass. A control question to check attention was included in position 21 (around half way through the question) in the four conditions.

The questionnaires were administered through the Moodle web platform. Participation was voluntary. The instructions were written at the start of the questionnaire and were read aloud by one of the researchers. Any doubts were answered and the time was limited to 40 minutes. The second pass was done after four weeks.

2.4. Data analysis

We used the Delphi model (Mokkink et al., 2010) to study the content validity. In stage 1, the initial questions were developed starting from the categories validated by Muedra (2020): morphology, phonetics, phonology and spelling, lexical-semantic level, syntax, text typology, oral and written expression processes/pragmatics, oral and written comprehension processes, literary resources.

To define them, we used the most recent bridge document of the autonomous region in which the study was performed, aimed at facilitating classroom planning (CEFIRE, 2015). The researchers extracted knowledge indicators and classified them independently within each category. A knowledge indicator is defined as a unit of knowledge expressed in a way that is specific and objective and, where applicable, is translatable to a behaviour that can be evaluated (Alfaro-Carvajal et al., 2022): for example, the indicator "Nouns. Classes: proper and common, individual and collective, concrete and abstract" was classified in the "morphology" category. In competence-based curricula, the selection of indicators seeks to facilitate the subsequent development of the instruments and means for evaluating the proposed competences.

To evaluate the functioning of these categories and of their indicators, activities from two collections of primary-school Spanish-language textbooks were classified. We observed that these categories included all of the indicators, and so they were used as a reference for defining the questions on the questionnaire.

In stage 2, we designed a bank of 142 questions, with a mean of 15 questions per category. We decided to create a multiple-choice questionnaire with four answer options, in line with studies that consider that distractors are functional if there are between three and five options (Downing, 2006; Haladyna, 2004; Haladyna & Downing, 1993). So, following the line of these studies, each question had one correct answer, another clearly incorrect one, and two that were incorrect but which aimed to induce mistakes. To prepare the questions, we chose activities from textbooks from four collections used widely throughout Spain and from specialist literature on didactics of language (Prado, 2004; Mendoza, 2003).

The researchers selected 40 questions from this initial bank considering the criteria of representativeness and presence in the curriculum. The resulting distribution of questions by category in the questionnaire was as follows: morphology (items 1, 2, 3, 4); phonetics, phonology, and spelling (items 5, 6, 7, 8); syntax (items 9, 10, 11, 12); linguistic and sociocultural variety (items 13, 14); lexical-semantic level (items 15, 16, 17, 18); literary resources (items 19, 20, 22, 23); text typology (items 24, 25, 26, 27); oral and written expression processes/pragmatics (items 28, 29, 30, 31, 32, 33, 34, 35, 36, 37); oral and written comprehension processes (items 38, 39, 40, 41). A control question to check attention was also included (item 21). Examples of possible questions and answers for each category can be found in the appendix. The scores for each item were 0 (incorrect option) and 1 (correct option); the scores for each item and for the questionnaire as a whole were obtained by calculating the mean value of the items involved.

In stage 3, this questionnaire was administered to a sample of 53 second-year students. They were asked about the intelligibility and the difficulty of the questions and answers, as well as of the control question. With the data from this sample, an analysis of the discriminatory capacity of the items and their difficulty was performed (Hurtado, 2018).

In stage 4, with the modified 40-question questionnaire, a first pass with the study sample was done to evaluate the construct validity and a second pass to check its reliability. Both studies were done using the scores from the items, the mean of the items from the final categories, and the total mean for the items by year (Table 1). All of the mean scores calculated in the study were normally distributed.

With the resulting items, we carried out a construct validity study using factor analysis in steps according to the figures for Cronbach's alpha (Taber, 2018), eliminating and averaging the items that the model indicated. For the analysis of construct validity, the average values of the variables collected in the first pass were used, grouped according to the results of the model. This analysis was also applied to the groupings of students by year to evaluate the suitability of the groupings for these variables.

To study reliability, we carried out a psychometric analysis of the variables taken at two different times (T1 and T2) using the test–retest method. We calculated the difference between the scores and the standard deviation of the difference; we applied the intraclass correlation coefficient (ICC) (Shrout & Fleiss, 1979) to the average values of each topic at the different times (T1 and T2) with the confidence intervals (95%), as well as the standard error of measurement, the repeatability coefficient, and the minimum detectable change (Beckerman et al., 2001; Bland & Altman, 1986). The ICC values were evaluated in line with the indications of previous studies (Landis & Koch, 1977).

We used the Bland–Altman plot to study the measurement error (Bland & Altman, 1996). To examine the mean error of the difference, we calculated the limits of agreement (95%) and their confidence intervals (Bland & Altman, 2010). To establish whether the error values between the passes were significant, we used *t* test for one sample on the differences in the T1 and T2 averages.

We calculated the development of the measurement error in relation to the average T1 and T2 values using a regression analysis (Bland & Altman, 1986). The floor/ceiling effect of the scores was calculated by comparing the percentages of participants with first and last quartile values for the scores from the first pass. If more than 15% of the study population was in one of these quartiles, the floor or ceiling effect was deemed to be present in the use of this tool (Terwee et al., 2007).

In order to rule out the possibility of an effect of the sex variable on the study subjects' scores, we applied a repeated means analysis of variance to the T1 and T2 scores with the analysis of the gender factor in time (T1–T2).

The difficulty of the questionnaire and of the items was calculated using the percentage scores of the sample compared to the total value. This value was also analysed by years.

To study the sample's knowledge of language, we calculated the means and deviations of all of the subjects by year, for each category and for the total scores. We used a one-factor ANOVA (year) to analyse the effect of the different years on the scores for the categories and the total score. Finally, we used the Bonferroni test for the post hoc contrasts.

3. Results

3.1. Content validity

Following stage 1's consultation of the specialist literature, the bridge document, and the categories established by Muedra (2020), in STAGE 2, a total of 142 questions were drawn up, of which 40 were sent to the experts after screening.

These experts unanimously determined that no essential content was missing and they stated that the questionnaire did evaluate the basic knowledge that a primary-school teacher should possess. They made suggestions regarding the wording of some items to reduce ambiguity or adjust the level of difficulty. Specifically, they proposed to increase the level of difficulty of some incorrect answers.

As a result, we modified 18 items. To ensure the comprehensibility and pertinence of the changes, two experts were asked to evaluate the questionnaire again. Four items were modified relating to the wording and level of difficulty of the answers.

As for the data deriving from the pilot test (stage 3), the students explained that the questionnaire in general seemed precise and intelligible to them. They did not identify the control question as such because they thought it was part of the content; as a result, we replaced it.

3.2. Construct validity and reliability

In stage 4, the analysis of the discriminatory capacity of the 40 items from the study sample, 24 had a low index of discrimination (<.125), 14 presented a very low index of difficulty (>93% correct answers), and 2 a very high index of difficulty (<10% correct answers). We eliminated the items that fulfilled both conditions (low capacity for discrimination and very high/low difficulty). This eliminated 10, leaving a questionnaire with 30 questions.

The factor analysis by steps in the study of the scale as a single factor pointed out the lack of consistency of these 30 items. The model indicated which items reduced the internal consistency and had to be eliminated. We grouped the remaining items into three categories: the MORF_LEX_SINT category comprised 3 items from Morphology, 3 from the lexical-semantic level, and 3 from syntax; the FFO_RECLIT category comprised 3 items from phonetics and phonology and 3 from literary resources; and the TT_PROEX_PRO-

COM category comprised 3 items from text typology, 6 from oral and written expression processes, and 4 from oral and written comprehension processes. With the averages of these 28 items grouped into three categories, the instrument achieved good internal consistency (Cronbach's alpha = .74).

The figure for consistency by gender was .71 (male) and .76 (female). The study by groups indicated an index of .75 for first year students, .75 for second year students, .67 for fourth year students, and .65 for master's students.

The mean values of the scores for both moments had a value slightly greater than the median of the scale (Table 1). The scores improved in T2 in general.

The mean error for the total scores was very small (.03), and the SEM (.04) displayed a low measurement error, with slightly higher values than the differences of means and lower than the SD of the difference. This happened in the same way in the groupings of the items. The RC also behaved well, giving values equal to or lower than two times the SD of the difference. The MDC indicated very limited sensitivity values for the instrument and showed real changes in the use of the instrument from values of 0.12 points in the total score.

Table 1. Test–retest values for the scores of the questionnaire (*n* = 190).

	M T1 (±SD)	M T2 (±DS)	M T1_T2 (±DS)	Dif. M T2 - T1 (±DS)	R	ICC (CI.95%)	RC	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_PRO-EX_PRO-COM	.64(.16)	.67(.16)	.65(.14)	.03(.15)*	.53*	.69 (.59-.77)	.30	.09	.24

Note: M = mean, T1 = time 1, T2 = time 2, SD = standard deviation, R = coefficient of correlation, ICC = intraclass correlation coefficient, CI = Confidence Interval, RC = repeatability coefficient, SEM = standard error of measurement, MDC = minimum detectable change; significant difference: *p <.05; **p <.01.

The strong intraclass correlation coefficients for the total test–retest scores (Table 1) indicated excellent reliability of the measures over time. However, significant differences were observed between the measurements from the two passes in the total scores and in two of the three groups of items.

Figures 1 and 2 show the absolute and relative values of the differences of the scores by their mean values. The mean value of the differences was .03 (SD .10) (Figure 1), equivalent to a percentage of error of 3.68% (Figure 2), which does not exceed the 5% acceptable probability of error. The regression analysis showed that the differences between the test and retest did not change as the means of the scores of the two times changed ($F_{(1,189)}$ = .2; *p* = .656; beta = .03). This indicated that the differences between the T1 and T2 scores did not vary in the different levels of knowledge of the sample.

The mean time that the sample took to answer the questionnaire was 14.67 minutes (SD 4.06).

No floor/ceiling effect was observed in the average scores obtained by the participants in the use of this questionnaire. No subject had average scores below .34 or above .89. However, 23% of subjects scored in the last quartile.

This study of the measurement error was also applied to the sample grouped by gender and year. We observed that the year groups with measurement error below 5% were the fourth year (0.8%) and master's (2.04%), while the percentage error for first years was 5.89%, and for second years, 7.54%.

FIGURE 1. Bland-Altman plot of absolute values of the scores.

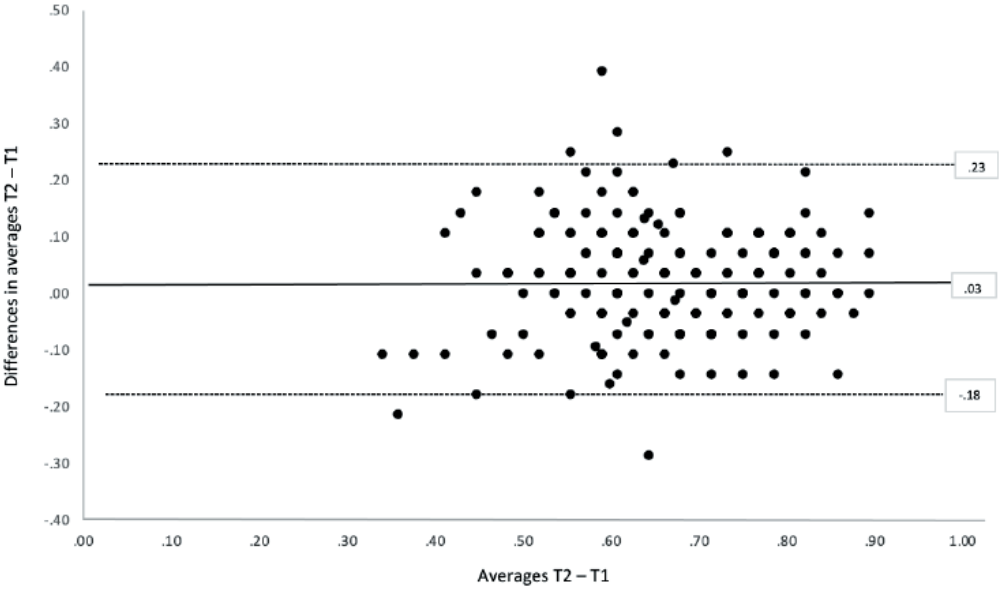
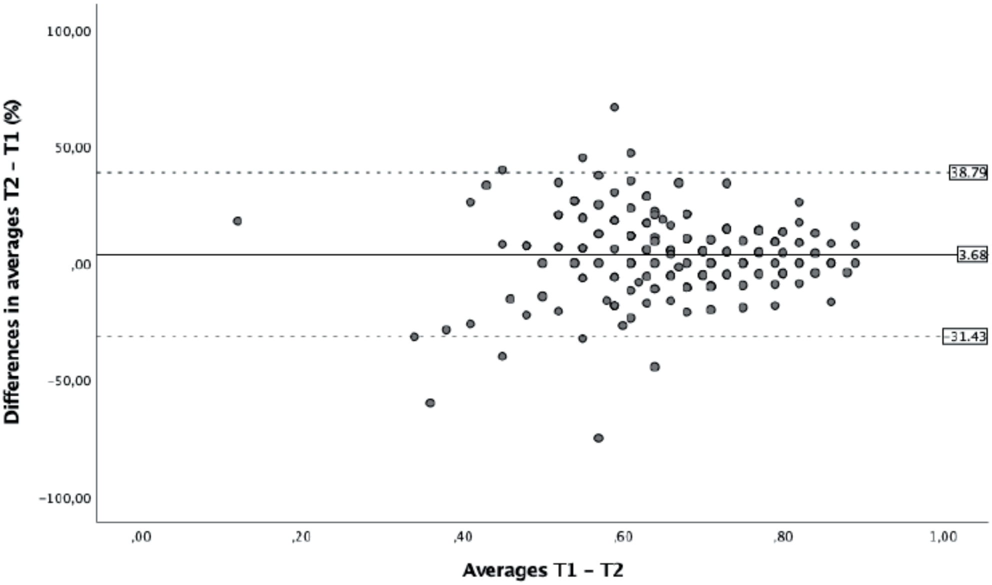


FIGURE 2. Bland-Altman plot of relative values of the scores.



With regards to gender, the results indicated a percentage error of 4.58% in men and 5.98% in women.

The repeated measures ANOVA indicated that gender did not influence the changes that occurred over time between the two measures ($F_{(1,126)} = 2.80$, $p = 598$).

The level of difficulty of the questionnaire for the study sample was medium. 17.86% did not pass the test, 50% had a score of between 5 and 7, and 32.14% exceeded the score of 7. These percentages varied between the different groups. The largest percentage of students who did not pass the test was in the fourth-year group (24.14%). The group with the highest percentage of students in the highest levels of scores (>7) was the master's group (60%).

3.3. Level of Spanish-language knowledge of pre-service teachers

Table 1 shows that the average level of knowledge was 0.64 points, which indicates that the sample studied achieved a medium score for knowledge in both the total score and in the different categories. Table 2 shows the results for the sample separated by year.

The study of frequencies of correct and incorrect answers shows that 12.6% of respondents failed the test (<5), 49.8% scored between 5 and 7, and 37.2% achieved a good score (>7).

As for the analysis of the effect of the different years, the results indicate that the scores were only different for the years variable in the total mean ($F_{(3,189)} = 3.408$; $p = .019$; eta squared = .045) and for the category FF_RECLIT ($F_{(3,189)} = 2.902$;

p = .036; eta squared = .052). The post hoc analysis indicated that there were differences in the total score between first-year and master's (diff. -.092; *p* = .022) and fourth-year and master's (diff. -.094; *p* = .017), and, for the FF_RECLIT category, between fourth-year and master's (diff. -.17; *p* = .022).

TABLE 2. Average total scores and scores by categories separated by year.

Category	Year	*N*	Mean	*SD*
MORF_LEX_SINT_01	First	65	0.658	0.169
	Second	40	0.671	0.180
	Fourth	65	0.634	0.165
	Master's	20	0.743	0.191
FF_RECLIT_01	First	65	0.633	0.213
	Second	40	0.625	0.238
	Fourth	65	0.595	0.228
	Master's	20	0.762	0.199
TT_PROEX_PROCOM_01	First	65	0.615	0.169
	Second	40	0.652	0.154
	Fourth	65	0.645	0.152
	Master's	20	0.694	0.134
Total_01	First	65	0.632	0.128
	Second	40	0.652	0.132
	Fourth	65	0.630	0.114
	Master's	20	0.724	0.104

Note: SD = standard deviation.

4. Discussion

The aim of this work was to design and validate an instrument to make possible the evaluation of Spanish-language CK of pre-service teachers. In line with research in other disciplines (Verdugo et al., 2019; Vásquez & Alsina, 2015; Godino et al., 2016), it started by considering specialist literature, the regulatory framework, and textbooks to generate a bank of questions, which is then subjected to a process of content and construct validation.

The Delphi method was used to guarantee high levels of validity; to avoid its potential drawbacks, we scrupulously complied with its characteristics, the implementation of its stages, and the selection of experts (Cabero & Infante, 2014) as well as its use in the target population. The Delphi method is especially useful when designing and validating an instrument if there are no instruments that fit the needs of the research (Andrés et al., 2019). Although there are some instruments that measure the knowledge that primary teachers should have for teaching students how to read in the English-speaking world (Washburn et al., 2016; Binks-Cantrell et al., 2012), the present study contributes the first instrument that makes it possible to measure Spanish-language CK of future teachers. The resulting questionnaire, comprising 28 questions and a control question, is shown to be valid and reliable for this purpose.

Both the experts and the users were unanimous with regards to the pertinence and validity of the instrument's content, and we implemented the changes relating to the formulation of questions and answers that they suggested. The pass with the pilot group also served to carry out a first discriminant analysis.

As for construct validity, after the process of elimination and grouping of items, the instrument achieved good internal consistency, both overall and for different genders and years, with lower consistency in higher years. This indicates robust functioning of the questionnaire independently of gender and year.

The analysis of the reliability of the measurements was backed by the mean values of the scores in the test–retest, the strong association between the measurements at the different moments, and the narrow margin of the measurement error. We observed that the average values improved significantly in the retest in the FF_RECLIT and TT_PROEX_PROCOM groups (groups in which they were found to more unstable), but the regression analysis showed that the differences observed between the scores did not vary as the average values increased. Therefore, we can state that there was no distorting effect in the process between measurements owing to learning.

From a methodological perspective, these results support the claims of Bland and Altman (1996): reliability studies with analysis of correlation of items at different moments provide insufficient information about their stability over time. Studies of reliability require in-depth consideration of the analysis of the measurement error, through analysis of relative and absolute reliability to confirm the effect of time on the use of the instrument (Vaz et al., 2013). On the one hand, we know that the precision of the instrument in the measurement of knowledge cannot present measurement indexes comparable to the study of more objective variables, as is habitual in the use of tools that evaluate complex concepts such as the ones tackled in this study. Nonetheless, as the regression analysis indicates, it does maintain its level of precision independently of the students' level of knowledge. We should add to this result regarding the reliability of the tool that the average measurement error is very low (0.03), and is equivalent to an acceptable error. In other words, there is a non-significant percentage of probability of different measurements (3.68%). The probability of a significant difference between the measures did not reach 5%.

To the good measurement error results, we must add the good psychometric behaviour of the scale, especially in the total values (Table 1). Independently of the groupings of the items into three categories, the scale has been evaluated as a single-factor scale that refers in general to the knowledge of Spanish language of the pre-service teachers. The variability of the measurement of the instrument (SEM = 0.04) was similar to the relative measurement error (0.03). The absolute measurement error or RC indicated that variations in the mean greater than 0.20 points correspond to measures with a value that will exceed the theoretical absolute error of the instrument and could be considered to be true variations. Also that changes in the scores on the questionnaire equal to or greater than the MDC value (0.12) could be regarded as real changes in students' knowledge.

Although the reliability of the questionnaire is good, the data extracted from its use must be interpreted with caution as the instrument was found to be less reliable in time in the analysis of the subgroup of women and in the groups with the least experience (first and second year). In this sense, the study has analysed in depth the behaviour of the measurement error by sex and by the different levels of training of the students, and the test worked better in the fourth-year and master's groups as well as in the male group. These groups had particular situations that could justify these results. On the one hand, 80% of the sample were female, which might explain the greater dispersal of scores between the results of the test and retest and consequent greater measurement error. On the other, the differences between the groups with the most and least experience could be because students from higher years had a more consolidated level of knowledge, independently of whether this was greater or lesser.

In the usability study, the questionnaire showed itself to be user-friendly and useful for teacher training: the average completion time was around 15 minutes and it was not difficult to understand.

The results regarding the level of knowledge were that 12.6% failed (<5), 49.8% scored between 5 and 7, and 37.2% achieved a good score (>7). From a mathematical perspective, the distribution of the scores could be said to be an acceptable mean level of knowledge (6-7 points) with a mean score of .64 and 87% of students passing the test. However, we must recall that the test was designed using questions on basic knowledge, from primary-school books. Therefore, we would not expect almost 50% of the sample to score between 5 and 7, even though it is true that similar shortcomings have been highlighted in studies of other areas (Verdugo et al., 2019; Vásquez & Alsina, 2015; Depaepe et al., 2013). This study highlights that the sample has a clear deficiency in disciplinary content in Spanish language.

It is also worrying that almost 13% of these future professionals do not pass the test. Teachers' CK is closely related to students' learning, and so fulfilling the requirement to know content in order to be able to teach it (Friedrichsen et al., 2009) is a responsibility for the people who train the Primary Teaching students and for the public institutions involved.

Moreover, the fourth-year students getting the lowest scores was unexpected. Although the general tendency is for knowledge to increases in higher years, the significance levels indicate that the evolution of the knowledge is not significant; this could be because disciplinary subjects are primarily taught in the first two years and are replaced in the last two years by the specifically didactic ones. The difference with master's students are to be expected, given that these students have broader disciplinary training.

Regarding the limitations of this work, we should note that it has not been possible to analyse convergent or criterion validity as there are no comparable instruments. We consider the study sample in the validation process to be adequate; however, other samples with different cultural characteristics should also be used.

5. Conclusions

Following the content and reliability analysis, we can state that the instrument presented here is valid and reliable for measuring pre-service teachers' Spanish-language CK.

At first glance, the first data seem to indicate that students have an acceptable knowledge of language; however, if we recall that the questionnaire seeks to measure minimum required knowledge, it is striking that half of the sample does not obtain more than what would be equivalent to a high pass/good grade.

The next phase of this research will involve administering this instrument to large samples of the population to establish whether this is simply because of the size of the sample or instead reveals a worrying reality about the training of primary-school teachers, a hypothesis that seems to be backed by research in other areas. Instruments like this one can help define and remedy these possible defects by providing the people who train teachers with consistent data to guide their practices more effectively.

Appendix. Examples of questions from the final questionnaire

The correct answer is shown in italics.

MORF_03. State which of these sentences does NOT include a verb in the subjunctive:

a) Maybe Teresa and Silvia will arrive late to the game. [Quizá Teresa y Silvia lleguen tarde al partido.]

b) Hopefully, it will rain more this spring. [Ojalá que llueva más esta primavera.]

c) If you were more interested, you would find studying easier. [Si tuvieras más interés, estudiar te resultaría más fácil.]

d) *Felipe will take part in the race on Sunday with his father. [Felipe participará en la carrera el domingo con su padre.]*

FFO_05. From the point of view of spelling, which of these sentences is correct?

a) *Tell me what is happening to you today. [Dime qué te pasa hoy.]*

b) I don't know where Paquita lives. [No sé donde vive Paquita.]

c) I have forgotten when I have an appointment with the doctor. [He olvidado cuando tengo cita con el médico.]

d) I don't know when it stopped hurting. [No sé en que momento dejó de dolerme.]

LEX_17. Choose the option in which all of the words are derived:

a) *Imperial, combative, volcanic, mountainous. [Imperial, combativo, volcánico, montañoso.]*

b) Combative, volcanic, love, lemon. [Combativo, volcánico, amor, limón.]

c) Volcanic, mountainous, table, heart. [Volcánico, montañoso, mesa, corazón.]

d) Love, lemon, table, heart. [Amor, limón, mesa, corazón.]

RECLIT_20. Choose the statement that is correct:

a) A sonnet has an assonant rhyme. [Un soneto tiene rima asonante.]

b) *A sonnet has 14 lines. [Un soneto tiene 14 versos.]*

c) A sonnet can be high or low art. [Un soneto puede ser de arte mayor o menor.]

d) A sonnet can have an unlimited number of stanzas. [Un soneto puede tener un número ilimitado de estrofas.]

TT_26. Identify the option that only contains oral genres:

a) *Dialogue, debate, press conference, and seminar. [El diálogo, el debate, la rueda de prensa y el coloquio.]*

b) Interview, presentation, recipe book, and news story. [La entrevista, la exposición, el recetario y la noticia.]

c) Personal diary, biography, travel book, and description. [El diario personal, la biografía, el libro de viajes y la descripción]

d) Dialogue, debate, personal diary, and seminar. [El diálogo, el debate, el diario personal y el coloquio.]

PROEX_31. Which of these statements does NOT correspond to planning writing:

a) Brainstorming. [Hacer una lluvia de ideas.]

b) *Correcting spelling*. *[Corregir la ortografía.]*

c) Looking for model texts. [Búsqueda de modelos.]

d) Outlining. [Hacer un esquema.]

References

Alfaro-Carvajal, C., Flores-Martínez, P., & Valverde-Soto, G. (2022). Conocimiento de profesores de matemáticas en formación inicial sobre la demostración: Aspectos lógico-matemáticos en la evaluación de argumentos [Knowledge of mathematics teachers in initial training regarding mathematical proofs: Logic-mathematical aspects in the evaluation of arguments]. *Uniciencia*, *36* (1), 140-165. https://doi.org/10.15359/ru.36-1.9

Andrés, I., Muñoz, M., Ruíz, G., Gil, B., Andrés, M., & Almaraz. A. (2019). Validación de un cuestionario sobre actitudes y práctica de actividad física y otros hábitos saludables mediante el método Delphi [Validation of a questionnaire on attitudes and practice of physical activity and other healthy habits through the Delphi method]. *Revista Española de Salud Pública*, *93*.

Beckerman, H., Roebroeck, M. E., Lankhorst, G. J., Becher, J. G., Bezemer, P. D., & Verbeek, A. (2001). Smallest real difference, a link between reproducibility and responsiveness. *Quality of Life Research, 10* (7), 571-578. https://doi.org/10.1023/A:1013138911638

Binks-Cantrell, E., Joshi, R. M., & Washburn, E. K. (2012). Validation of an instrument for assessing teacher knowledge of basic language constructs of literacy. *Annals of dyslexia*, *62*, 153-171. https://doi.org/10.1007/s11881-012-0070-8

Bland, J. M., & Altman, D. (1986). Statistical methods for assessing agreement between two methods of clinical measurement. *The Lancet, 327* (8476), 307-310. https://doi.org/10.1016/S0140-6736(86)90837-8

Bland, J. M., & Altman, D. G. (1996). Statistics notes: Measurement error. *Bmj, 312* (7047), 41-42. https://doi.org/10.1136/bmj.312.7047.1654

Bland, J. M., & Altman, D. G. (2010). Statistical methods for assessing agreement between two methods of clinical measurement. *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. *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. *Chemistry Education International*, *6* (1), 1-2.

Cabero, J., & Infante, A. (2014). Empleo del método Delphi y su empleo en la investigación en Comunicación y Educación [Using the Delphi method and its use in communication research and education]. *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-UnzX42B-P746IHFC-HytD/view?usp=sharing

Chetty, R., Friedman, J. N., & Rockoff, J. E. (2011). *The long-term impacts of teachers: Teacher value-added and student outcomes in adulthood*. National Bureau of Economic Research. https://doi.org/10.3386/w17699

Depaepe, F., Verschaffel, L., & Kelchtermans, G. (2013). Pedagogical content knowledge: A systematic review of the way in which the concept has pervaded mathematics educational research. *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. In S. M. Downing, & T. M. Haladyna (Eds.), *Handbook of test development* (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., & Volkmann, M. J. (2009). Does teaching experience matter? Examining biology teachers' prior knowledge for teaching in an alternative certification program. *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., & Stuhlsatz, M. A. (2019). Teacher pedagogical content knowledge, practice, and student achievement. *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., & Díaz-Batanero, C. (2016). Evaluación de conocimientos didáctico-matemáticos sobre visualización de objetos tridimensionales en futuros profesores de educación primaria [Assessing didactic-mathematical knowledge of prospective primary school teachers on visualization of three-dimensional objects]. *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*. Routledge. https://doi.org/10.4324/9780203825945

Haladyna, T. M., & Downing, S. M. (1993). How many options is enough for a multiple-choice test item? *Educational and Psychological Measurement*, *53* (4), 999-1010. https://doi.org/10.1177/0013164493053004013

Harris, D. N., & Sass, T. R. (2011). Teacher training, teacher quality and student achievement. *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 [Relationship between the difficulty and discrimination indices]. *Revista digital de investigación en docencia universitaria*, *12* (1), 273-300. http://dx.doi.org/10.19083/ridu.12.614

Kleichkmann, T., Richter, D., Kunter, M., Elsner, J., Besser, M., Krauss, S., & Baumert, J. (2013). Teachers' content knowledge and pedagogical content knowledge: The role of structural differences in teacher education. *Journal of teacher education*, *64* (1), 90-106. https://doi.org/10.1177/0022487112460398

Landis, J. R., & Koch, G. G. (1977). The measurement of observer agreement for categorical data. *Biometrics, 33,* (1), 159-174. https://doi.org/10.2307/2529310

Mendoza, A. (2003). *Didáctica de la lengua y la literatura para educación primaria [Didactics of language and literature for primary education]*. Pearson Educación.

Ministerio de Educación, Cultura y Deporte. (2012). *TEDS-M. Informe español. Estudio internacional sobre la formación inicial en matemáticas de los maestros [TEDS-M. Spanish report. International study on initial teacher training in mathematics].* 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., & 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. *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 [Analysis of oral competence in Spanish language textbooks for primary education]* [Master Thesis]. Universitat de València.

Opfer, V. D., & Pedder, D. (2011). Conceptualizing teacher professional learning. *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. *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 [Didactics of language and literature for education in the 21st century]*. La Muralla.

Shrout, P. E., & Fleiss, J. L. (1979). Intraclass correlations: Uses in assessing rater reliability. *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. *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. *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., & 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)*. 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., & De Vet, H. C. (2007). Quality criteria were proposed for measurement properties of health status questionnaires. *Journal of Clinical Epidemiology, 60* (1), 34-42. https://doi.org/10.1016/j.jclinepi.2006.03.012

Vásquez, C., & Alsina, A. (2015). Conocimiento didáctico-matemático del profesorado de educación primaria sobre probabilidad: diseño, construcción y validación de un instrumento de evaluación [Primary school teachers' didactic-mathematical knowledge when teaching probability: development and validation of an evaluation instrument]. *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., & Andreou, P. (2013). The case for using the repeatability coefficient when calculating test-retest reliability. *PloS One*, *8* (9), e73990. https://doi.org/10.1371/journal.pone.0073990

Verdugo, J. J., Solaz, J. J., & Sanjosé, V. (2019). Evaluación del conocimiento científico en maestros en formación inicial: el caso de la Comunidad Valenciana [Assessment of pre-service teachers' science knowledge: the case of Valencian Community in Spain]. *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., & Arrow, A. (2016). Preservice teacher knowledge of basic language constructs in Canada, England, New Zealand, and the USA. *Annals of dyslexia*, *66*, 7-26. https://doi.org/10.1007/s11881-015-0115-x

Authors' biographies

María Teresa Muedra-Pedris. Student on the Specific Didactics Doctoral Programme at the Universidad de Valencia. She is a member, along with the other authors, of the «Pedagogical content knowledge: foundations of teachers' didactic analysis and action» research group (CDC-GIUV2016-280).

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

Manuel Monfort-Pañego. Associate Professor in the Department of Physical, Artistic, and Music Education Teaching of the Universidad de Valencia. Doctor of Physical Education from the Universidad de Valencia. His research interests focus on primary-school teacher training, principally the development and validation of measurement instruments to assess students' knowledge and habits in relation to physical education. He is a member, along with the other authors, of the «Pedagogical content knowledge: foundations of teachers' didactic analysis and action» research group (CDC-GIUV2016-280).

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

Ángela Gómez-López. Associate Professor in the Department of Language and Literature Teaching of the Universidad de Valencia. Doctor of Specific Didactics from the Universidad de Valencia. Her research interests focus on primary-school teacher training: learning and teaching of foreign languages, reading comprehension and control of comprehension (metacognition) in L2 and didactic analysis of school content in first and second languages. She is a member, along with the other authors, of the «Ped-

agogical content knowledge: foundations of teachers' didactic analysis and action» research group (CDC-GIUV2016-280).

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

Eva Morón-Olivares. Associate Professor in the Department of Language and Literature Teaching of the Universidad de Valencia. Doctor of Hispanic Philology from the Universidad de Granada. Her research interests centre on training primary-school teachers, especially for teaching literature and the didactic analysis of school content in first and second languages. She is a member, along with the other authors, of the «Pedagogical content knowledge: foundations of teachers' didactic analysis and action» research group (CDC-GIUV2016-280).

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