



Development and content based validity evidence for the Social Thinking and Identity Diversity PSDI instrument for qualitative data collection in higher education

Desarrollo y evidencias de validez basadas en el contenido del instrumento pensamiento social y diversidad identitaria PSDI para la recogida de datos cualitativos en educación superior

Delfín Ortega-Sánchez¹

Carlos Pérez-González²

¹ Universidad de Burgos. Burgos, España.

² Universidad de Burgos. Burgos, España.

Recibido: 16-07-2026 | Revisado: 24-08-2026 | Aceptado: 26-09-2026 | Publicado: 29-09-2026

Autor de correspondencia: Delfín Ortega-Sánchez

ABSTRACT

The aim of this study is to develop and obtain content-based validity evidence for the Social Thinking and Identity Diversity (PSDI) instrument, a structured guide for individual interviews and focus groups designed to explore representations, experiences, and positionings related to social thinking, inclusive education, and identity diversity in higher education. Using a sequential instrumental design, 34 expert judges from three areas of expertise evaluated 19 stimuli across two successive rounds according to their relevance, representativeness, essentiality, and five technical quality criteria. The first round identified four stimuli that required reformulation. Following their revision, all 19 met the established decision rules (Aiken's $V = .912-.971$, CVC = $.934-.978$, and CVR = $.588-.941$). Agreement and interrater reliability increased substantially (Kendall's $W = .843$; Fleiss' $\kappa = .803$; Krippendorff's ordinal $\kappa = .837$), with no significant differences across areas of expertise. The results indicate that the PSDI constitutes a methodologically grounded and sufficiently refined qualitative guide for proceeding to pilot testing in university educational settings.

Keywords: Social thinking; identity diversity; inclusive education; content validity; higher education.

RESUMEN

El objetivo de este estudio es desarrollar y obtener evidencias de validez basadas en el contenido del instrumento 'Pensamiento social y diversidad identitaria' (PSDI), una guía estructurada para entrevistas individuales y grupos focales destinada a explorar representaciones, experiencias y posicionamientos relacionados con el pensamiento social, la educación inclusiva y la diversidad identitaria en educación superior. Mediante un diseño instrumental secuencial, 34 jueces expertos de tres ámbitos de especialización evaluaron 19 estímulos en dos rondas sucesivas según su relevancia, representatividad, esencialidad y cinco criterios de calidad técnica. La primera ronda identificó cuatro estímulos que requirieron reformulación. Tras su revisión, los 19 superaron las reglas de decisión establecidas (V de Aiken = $.912-.971$, CVC = $.934-.978$ y CVR = $.588-.941$). La concordancia y la fiabilidad interevaluador aumentaron sustancialmente (W de Kendall = $.843$; κ de Fleiss = $.803$; κ ordinal de Krippendorff = $.837$), sin diferencias significativas entre ámbitos de especialización. Los resultados obtenidos indican que el PSDI constituye una guía cualitativa metodológicamente fundamentada y suficientemente depurada para avanzar hacia su pilotaje en contextos formativos universitarios.

Palabras clave: Pensamiento social; diversidad identitaria; educación inclusiva; validez de contenido; educación superior.

INTRODUCTION

Identity diversity constitutes a structural condition of contemporary higher education and, at the same time, an educational challenge that transcends the mere coexistence of students with diverse characteristics. Available evidence indicates that experiences of interaction with diverse peers may be associated with more complex forms of reasoning and a greater disposition to engage in critical thinking; however, these effects depend on the nature and quality of such interactions and do not arise automatically from group heterogeneity. In this regard, Hurtado (2005) showed that certain experiences of interaction with diverse peers were associated with the development of cognitive and sociocognitive skills, whereas negative interactions were associated with a lower disposition toward critical thinking. In the Ibero-American context, Pérez Serrano and SarrateCapdevila (2013) likewise linked cultural diversity and inclusion in the university context to the construction of the foundations of democratic citizenship. From this perspective, diversity acquires educational value when it is operationalized through institutional and pedagogical conditions conducive to the recognition of difference, reflective interaction, and participation.

The most recent scientific literature makes it possible to organize the skills involved in this process into three interrelated domains, without presupposing that they constitute clearly defined independent psychometric dimensions. A first, cognitive-interpersonal domain comprises cognitive empathy, perspective-taking, and intercultural competence; a second, emotional-relational domain integrates self-regulation, communication, and the establishment of respectful relationships; and a third, critical-analytical domain encompasses critical thinking, problem-solving, and decision-making. In this regard, the study by Peifer and Taasobshirazi (2022), conducted with a large sample of higher education students, provides evidence of the relationships among ethnic identity, intercultural competence, and cognitive empathy. Likewise, the evaluation of a university multicultural psychology course showed pretest–posttest changes in cultural skills, problem-solving, leadership, attitudes toward social justice, and empathy, together with qualitative findings related to intercultural contact and perspective-taking (Simons et al., 2023). Complementarily, reviews on critical thinking and inclusive education highlight the relevance of critical-analytical and socioemotional competencies, while also noting considerable methodological heterogeneity and variability in the measurement procedures employed (López-Ruiz et al., 2021; Nevárez Jiménez et al., 2025). Social thinking is conceived as a cross-cutting construct that integrates cognitive-interpersonal, emotional-relational, and critical-analytical processes. Along these lines, inclusion and identity diversity constitute the contexts through which processes of perspective-taking, social interpretation, attitude appraisal, critical thinking, problem-solving, and decision-making are elicited (Calderón Jaramillo, 2018; Peifer & Taasobshirazi, 2022; Solbes & Torres Merchán, 2012).

The literature further agrees that the development of these skills requires complementary actions at different levels. At the institutional level, inclusive policies, curricular adaptation, and the creation of participatory environments establish the conditions necessary for diversity to become a meaningful educational experience. At the teaching level, initial and continuing teacher education, sensitivity to diversity, and the capacity to plan, mediate, and assess inclusive practices constitute factors of significant influence (Paz Maldonado, 2018). At the specifically didactic level, collaborative learning and reflective processes have been highlighted as strategies for promoting social skills (Duran & Parra, 2014); the analysis of relevant social problems or socially acute questions as a means of fostering the construction of contextualized conceptual frameworks and the development of decision-making and interaction skills (Calderón Jaramillo, 2018); and the treatment of socio-scientific issues as a resource for stimulating competencies associated with critical thinking (Solbes & Torres Merchán, 2012). Likewise, university extension activities have been proposed as suitable educational settings for systematically promoting teamwork and social responsibility (Ordaz Hernández, 2013).

This conceptual convergence contrasts with the fragmentation of the procedures employed to address this research problem. Indeed, the available studies combine qualitative approaches, descriptive designs, evaluations of educational interventions, systematic reviews, and theoretical-documentary analyses, with substantial differences in the types of diversity considered, the populations analyzed, the outcomes examined, and the levels of intervention. Even when quantitative indicators are incorporated, they are usually confined to specific constructs. Thus, for example, the descriptive study by Betancourth-Zambrano et al. (2020) assessed critical thinking in 119 Social Work students without providing an integrated measure of the relationships between social thinking and identity diversity. This heterogeneity hinders direct comparison across studies and highlights the value of having sufficiently structured instruments and qualitative data-collection procedures that make it possible to examine, within a common framework, conceptualizations

of inclusion, educational experiences, professional practices, attitudes, and training needs related to identity diversity.

In this context, the present study focuses on the development and collection of content-based validity evidence for Social Thinking and Identity Diversity (PSDI), an instrument conceived as a structured guide for conducting individual interviews and focus groups in higher education. Its purpose is to provide an organized set of prompts designed to qualitatively explore representations, experiences, and positionings related to social thinking, inclusive education, and identity diversity, without seeking to generate psychometric scores or diagnose individual competencies. Specifically, this study addresses the following objectives: (a) to examine, through expert judgment, the relevance, representativeness, and essentiality of the proposed prompts; (b) to refine formulations that do not satisfactorily meet the *a priori* criteria; (c) to obtain complementary evidence regarding their clarity, coherence, absence of bias, linguistic appropriateness, and precision; (d) to estimate the agreement and reliability of the ratings using coefficients appropriate for a multi-rater design with ordinal data; and (e) to determine whether technical ratings differ systematically across the areas of expertise represented on the expert panel.

METHODS

The study was designed as instrumental research aimed at developing the Social Thinking and Identity Diversity (PSDI) instrument and obtaining content-based validity evidence and evidence of inter-rater reliability (American Educational Research Association et al., 2014; Ato et al., 2013).

A panel of 34 expert judges was constituted and selected through purposive sampling on the basis of their suitability for the conceptual and methodological domain of the instrument. The panel was structured into three complementary areas of expertise: Social Sciences Education, social thinking, and citizenship education ($n = 12$); inclusive education ($n = 11$); and educational research methods ($n = 11$). Eligibility criteria required experts to hold a postgraduate qualification related to at least one of the domains under evaluation, to have a minimum of five years of teaching, research, or professional experience, and to demonstrate an established track record in research, scientific publications, or participation in projects directly related to the content being evaluated. The descriptive profile of the panel comprised 18 women (52.9%) and 16 men (47.1%), with a mean professional experience of 16.1 years ($SD = 6.4$; range = 6–31), and affiliations with 21 institutions across seven Ibero-American countries (Spain, $n = 16$; Mexico, $n = 5$; Portugal, $n = 4$; Chile, $n = 3$; Argentina, $n = 2$; Colombia, $n = 2$; Ecuador, $n = 2$). By type of institutional affiliation, 23 judges were affiliated with public universities, seven with private universities, and four with research or teacher education centers. All 34 expert judges participated in both rounds, with no attrition, and individually rated the 19 prompts on both occasions. The evaluations were conducted independently, with no exchange of scores or communication of aggregate results between rounds.

The PSDI is a structured guide for collecting qualitative information through individual interviews or focus groups on social thinking, inclusive education, and identity diversity in higher education contexts. The version submitted for evaluation comprised 13 main questions and 6 probing questions, corresponding to 19 evaluable prompts organized into five blocks: (a) Representation of the concept of inclusive education; (b) Definition of inclusive education and educational experience in identity diversity, subdivided into B-1 and B-2; (c) Professional experience in identity diversity; (d) Education for identity diversity in higher education; and (e) Education for identity diversity in teaching practice. The correspondence between the construct, its domains, and the prompts is specified in Table 1. The domains represent partially overlapping areas of conceptual coverage rather than independent psychometric dimensions.

To obtain complementary evidence regarding content quality, an expert-judgment form comprising three rating systems was designed (Table 2 and Figure 1). First, each prompt was rated in terms of its content relevance and representativeness using a four-category ordinal scale ranging from 1 (low adequacy) to 4 (maximum adequacy), intended for the calculation of Aiken's V and, complementarily, the content validity coefficient (CVC) (Aiken, 1985; Hernández-Nieto, 2002). Second, each prompt was classified as essential, useful but not essential, or not necessary, in accordance with Lawshe's (1975) procedure, to obtain the content validity ratio (CVR). Third, an independent technical evaluation was conducted, also using a 1-to-4 scale, for five criteria adapted from the procedure employed in the validation of the DICHIE instrument (Ortega-Sánchez & Bernal López-Sanvicente, 2023): clarity of wording, internal coherence, absence of bias or response inducement, linguistic appropriateness, and precision. For operational purposes, relevance was defined as the

degree to which the prompt was substantively important for the purpose of the evaluation, and representativeness as the degree to which it adequately captured the content of the theoretical block to which it was assigned; both aspects were integrated into the four-level ordinal judgment of substantive adequacy (1 = low, 2 = insufficient, 3 = moderate, and 4 = high adequacy) used to calculate Aiken's V and the CVC. For the calculation of the CVR, the three-category judgment was dichotomously recoded in accordance with Lawshe: essential = 1; useful but not essential or not necessary = 0; only the essential category contributed to n_e .

For the five technical dimensions, 1 indicated failure to meet the criterion, 2 need for substantial revision, 3 adequacy with possible minor adjustments, and 4 full satisfaction of the criterion. The form also included an open-ended field in which the judges could justify their ratings and propose semantic or wording modifications. The five technical criteria were operationally defined as follows: clarity, comprehensible and unambiguous wording; internal coherence, logical congruence of the prompt with the block and the evaluative purpose; absence of bias, absence of response inducement, presupposition, or direction; linguistic appropriateness, correspondence of register and terminology with the target university population; and precision, direct correspondence between what is asked and the content intended to be elicited.

Table 1. Specification of the representation of social thinking in the PSDI

Block	Prompts	Function related to social thinking	Predominant domain
A	1-3	Social representation, consequences, and interpretation of attitudes.	Cognitive-interpersonal, emotional-relational, and critical-analytical.
B-1	4, 5, 5.1	Conceptualization and relationship between inclusion and social thinking.	Cognitive-interpersonal and critical-analytical.
B-2	6, 6.1	Background and educational experiences.	Cross-cutting contextual
C	7, 7.1, 8, 9	Professional experience, inclusive action, and assessment of needs.	Cognitive-interpersonal and critical-analytical.
D	10-11.1	Interpretation of attitudes and positionings, problem-solving, and educational proposals.	Cognitive-interpersonal, emotional-relational, and critical-analytical.
E	12-13	Assessment of feasibility and prioritization.	Critical-analytical

Note. The classification identifies the predominant function of each block and does not imply subscales or latent dimensions. Prompts 6 and 6.1 primarily serve a contextual function.

Figure 1. Relationship between the rating systems, recoding procedures, and resulting indices

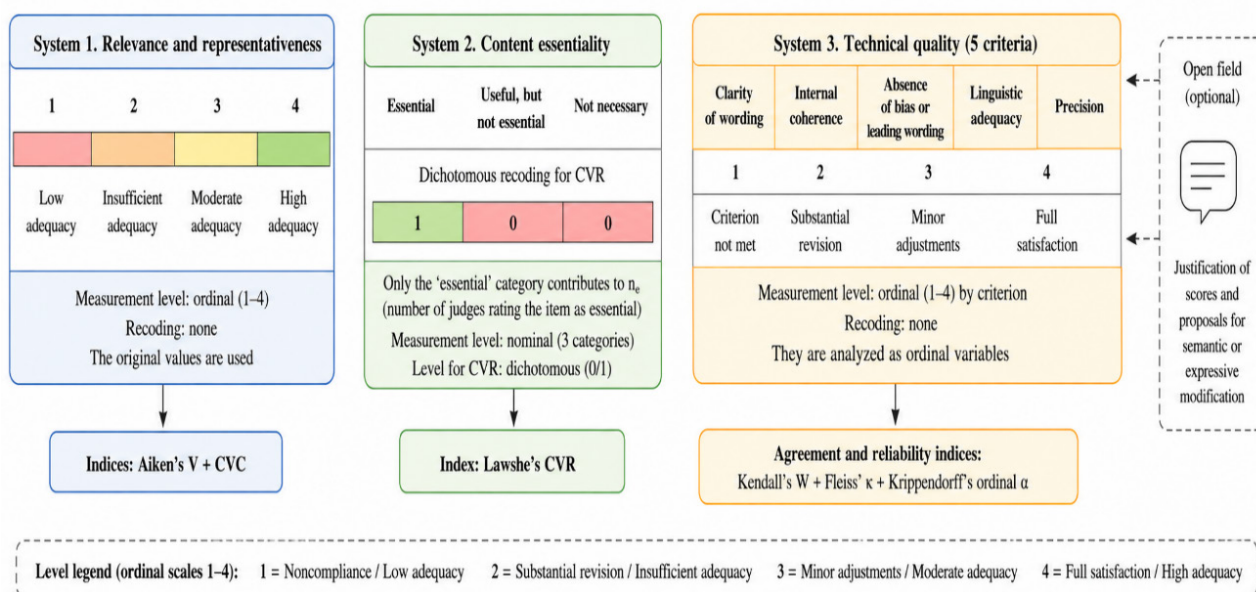


Table 2. Expert-judgment rating systems, level of measurement, recoding, and linkage to validity indices

Rating system	Purpose	Original scale	Level of measurement	Recoding for the index	Use and index obtained	Methodological source
1. Content relevance and representativeness	To assess the extent to which the question is substantively important (relevance) and adequately captures the content of the theoretical section (representativeness).	4-category ordinal scale (substantive adequacy): 1 = Low 2 = Insufficient 3 = Moderate 4 = High	Ordinal (1-4)	No recoding. Original values are used.	Calculation of: Aiken's V (for each question) Content Validity Coefficient (CVC) (complementary analysis)	Aiken (1985); Hernández-Nieto (2002)
2. Content essentiality	To determine whether the question is essential for the assessment purpose.	3-category nominal scale: Essential Useful but not essential Not necessary	Nominal (3 categories)	Dichotomous recoding for CVR: Essential = 1 Useful but not essential = 0 Not necessary = 0 Only the essential category contributes to n_e (number of judges rating the question as essential).	Calculation of: Lawshe's Content Validity Ratio (CVR) (for each question)	Lawshe (1975)
3. Technical quality of the question (5 criteria)	To independently assess the technical quality of the question according to five criteria adapted from the DICHIE instrument.	4-category ordinal scale (technical adequacy): 1 = Failure to meet the criterion 2 = Substantial revision 3 = Minor adjustments 4 = Full satisfaction	Ordinal (1-4)	No recoding. The criteria are analyzed as ordinal variables.	Calculation of inter-rater agreement for each criterion using: Kendall's κ W Fleiss' κ Krippendorff's ordinal α (reliability analysis)	Kendall & Babington Smith (1939); Fleiss (1971); Krippendorff (2018); Ortega-Sánchez & Bernal López-Sanvicente (2023)

Note. CVC = Content Validity Coefficient; CVR = Content Validity Ratio; V = Aiken's V; κ = kappa; α = alpha.

The process followed a sequential instrumental design comprising two rounds of expert judgment. In the first phase, the theoretical correspondence among the purpose of the PSDI, its blocks, and the 19 interview prompts was reviewed. Subsequently, ratings were obtained from the 34 expert judges for the first version of the instrument, while maintaining independence among records. The ratings were analyzed jointly with the open-ended comments in order to identify problems related to relevance, representativeness, clarity, coherence, bias, linguistic appropriateness, or precision.

Prior to data analysis, the decision rules were established. For Aiken's V, a specific conservative criterion of $V \geq .80$ and a lower bound of the 95% CI $\geq .70$ was adopted; the intervals were estimated using the score-based procedure described by Penfield and Giacobbi (2004). This criterion was not intended as a universal cutoff, but rather as an a priori decision rule. The CVC was used solely as a complementary sensitivity analysis, with .80 considered an operational reference value (Hernández-Nieto, 2002). For the CVR, the decision rule was applied to the integer count of essentiality judgments. With $N = 34$, at least 23 ratings of essential were required, equivalent to a CVR of .352941, reported as .353, in accordance with the exact binomial procedure revised by Ayre and Scally (2014). Finally, a $Mdn \geq 3$ was required for each of the five technical criteria.

The joint application of these rules identified four prompts that required reformulation (5, 6.1, 10.2, and 11.1). The modifications were restricted to the identified problems and preserved the intended conceptual domain. In the second round, each judge again received the 19 prompts, four reformulated and 15 unchanged, and rated them using the same procedure, without receiving aggregate information on the previous results.

The study involved the voluntary participation of expert judges, whose ratings were collected exclusively for the purposes of evaluation and of obtaining content-based validity evidence for the instrument. The research was conducted in accordance with the ethical principles applicable to research involving human participants and, specifically, with the ethical standards for educational research (American Educational Research Association, 2011; British Educational Research Association, 2024), as well as with the relevant data protection principles established in Regulation (EU) 2016/679 (European Parliament & Council of the European Union, 2016).

The personal and professional information collected was limited to that strictly necessary to characterize the suitability of the expert panel; no sensitive, clinical, or directly identifying data were collected. The ratings were treated confidentially, and the results were analyzed and reported in aggregate form, without individually identifying the participants. The processing and storage of the information were carried out in accordance with the principles of data minimization, purpose limitation, confidentiality, and security. Having expressly identified the competent body and the applicable regulatory basis within the university institution with which the authors are affiliated and under whose framework the present research was conducted, the submission of a formal request for an ethical opinion was not required, and no case or reference number was generated. Nevertheless, prior to participation, the expert judges received sufficient information regarding the objectives and procedures of the study, the voluntary nature of their participation, the intended processing of their ratings, and the guarantees of confidentiality, and provided prior informed consent.

Content-based evidence was examined using complementary indicators, avoiding the interpretation of any single coefficient as sufficient evidence of validity. Aiken's V was calculated for each prompt on the basis of the ordinal relevance-representativeness ratings and was accompanied by 95% CIs (Aiken, 1985; Penfield & Giacobbi, 2004). As a sensitivity check, Hernández-Nieto's (2002) CVC was estimated. Content essentiality was analyzed using Lawshe's CVR, $CVR = (n_e - N/2) / (N/2)$, where n_e represents the number of judges who considered the prompt essential and N the total number of judges (Ayre & Scally, 2014; Lawshe, 1975). Additionally, the mean CVR across the 19 prompts was reported as a descriptive global index.

Given the ordinal nature of the five technical variables and the participation of 34 judges, agreement was examined using several coefficients with distinct statistical functions. The overall Kendall's W was calculated across 95 evaluation units, corresponding to the 19 prompts $\times$ five technical criteria. For each judge, these ratings were transformed into ranks, with average ranks assigned to ties and the corresponding tie correction applied to the denominator of W, whereas criterion-specific analyses were conducted on matrices of 19 units $\times$ 34 judges (Kendall & Babington Smith, 1939). The overall W is interpreted as a descriptive summary of agreement across the full set of technical ratings, rather than as the mean of the five criterion-specific W coefficients or as evidence of unidimensionality. Fleiss' κ , in turn, was estimated as a chance-corrected nominal multi-rater measure, whereas Krippendorff's α with an ordinal metric was used as the principal indicator of inter-rater reliability because it incorporates the ordering of the categories (Fleiss, 1971; Hayes & Krippendorff, 2007; Krippendorff, 2018).

As a complementary check of inter-rater agreement and of the sensitivity of the content-based validity evidence obtained, Cohen's κ coefficients were calculated for each of the possible pairs of judges:

$$\binom{34}{2} = \frac{34!}{2!(34-2)!} + \frac{34 \times 33}{2} = 561$$

Because Cohen's κ is defined to assess agreement between two raters, the 561 coefficients obtained were summarized descriptively using their median and minimum–maximum range, without interpreting them as a global estimate of multi-rater agreement or as statistically independent coefficients. This procedure was used exclusively to examine variability in pairwise agreement and as a sensitivity analysis complementary to the main content-validity indices (Cohen, 1960).

Similarly, Bangdiwala's agreement chart and the BN coefficients (unweighted strict agreement coefficient) and weighted B (, weighted agreement coefficient) were calculated pairwise within each of the three professional strata (176 pairwise comparisons at the descriptive level): 66 in the Social Sciences Education group, 55 in inclusive education, and 55 in education and research. This operationalization respected the paired nature of Bangdiwala's statistic and the ordering of the categories (Bangdiwala & Shankar, 2013; Muñoz & Bangdiwala, 1997). In the pairwise analyses, Cohen's κ was calculated without weighting on the same four categories treated nominally, whereas $B_{N \times N}$ quantified exact agreement and weighted B incorporated partial agreement between ordinally adjacent categories using quadratic weights, while preserving the order $1 < 2 < 3 < 4$.

To examine potential differences among the three areas of expertise, the Kruskal–Wallis H test was applied (Kruskal & Wallis, 1952). For each technical criterion, the mean of the 19 ratings was calculated for each judge as an aggregated descriptive index of the 1–4 coding. This summary

assumes such numerical coding without attributing interval-scale properties to the criterion; the Kruskal–Wallis test was subsequently applied to the ranks of the 34 individual indices distributed across the three groups (Harpe, 2015). The magnitude of the differences was estimated using $\varepsilon^2_H = (H - k + 1) / (N - k)$, with $k = 3$ and $N = 34$, setting negative estimates to 0. Post hoc comparisons were reserved for cases in which the omnibus test reached statistical significance. The coded assignment of the 34 judges to the three areas of expertise is presented in Table S2 of the supplementary material.

No coefficient was interpreted in isolation or by means of rigid cutoff points or qualitative labels. Interpretation was based on an integrated assessment of the magnitude of the estimator, the precision reflected by its confidence intervals, the level of measurement of the data, and the specific statistical function for which each coefficient was calculated.

RESULTS

The first round showed an overall favorable evaluation, although the results were not homogeneous across the 19 prompts. Aiken's V ranged from .686 to .931 ($M = .855$), and the lower bounds of the corresponding 95% CIs ranged from .591 to .865. The CVC yielded values between .765 and .949 ($M = .891$). CVR values, in turn, ranged from .176 to .824, with an overall mean of .508.

The simultaneous application of the a priori criteria identified four prompts with insufficient evidence in the first round (Table 3 and Table S1). In all of these cases, Aiken's V was below .80 and/or the lower confidence bound fell below .70; likewise, fewer than 23 of the 34 judges selected the essential category, and therefore their CVR values did not reach the critical value of .353.

Table 3. Content validity evidence for the reformulated prompts across the two rounds of expert judgment

Estímulo	Vr1	95 % CI for V	CVRr1	Main modification	Vr2	CVRr2
5	.686	[.591, .768]	.235	Neutralization of the assumption of a direct relationship between inclusion and social thinking.	.951	.824
6.1	.725	[.632, .803]	.294	Greater specification of the type of educational experience being requested.	.912	.588
10.2	.725	[.632, .803]	.176	Reduction of evaluative inducement when asking about other teachers.	.912	.588
11.1	.755	[.663, .828]	.294	Disaggregation and greater specification of the components of teacher education.	.931	.706

Note. Vr1-r2 = Aiken's V (first and second rounds); 95% CI = 95% confidence interval; CVRr1-r2 = content validity ratio (first and second rounds).

The 15 unchanged prompts likewise showed increases between rounds (V : $M = .890 \rightarrow .942$, $\Delta M = .052$; CVR : $M = .576 \rightarrow .769$, $\Delta M = .192$), whereas the changes were greater for the four reformulated prompts (V : $M = .723 \rightarrow .926$, $\Delta M = .203$; CVR : $M = .250 \rightarrow .676$, $\Delta M = .426$). Therefore, this pattern does not support attributing the overall improvement solely to the textual reformulation undertaken.

In accordance with the results obtained in the first phase and in response to the suggested modifications, the four critical prompts were reformulated as follows: 5. "What relationship, if any, do you identify between the definition of inclusive education you have just formulated and the development of social thinking skills?"; 6.1. "If you have addressed this issue, describe a specific experience and indicate the context, content, methodology, and type of activity through which identity diversity was addressed"; 10.2. "Based on your experience, what responses or positions do you observe among university faculty regarding education for identity diversity?"; and 11.1. "What content, educational experiences, and teaching strategies do you consider priorities in initial teacher education for addressing identity diversity?".

Following the introduction of these modifications, all 19 prompts jointly met the predefined criteria for retention. In the second round, Aiken's V ranged from .912 to .971 ($M = .939$), with lower bounds of the 95% CI ranging from .841 to .917. The CVC, in turn, ranged from .934 to .978 ($M = .954$), whereas all CVR values exceeded the critical value for 34 judges, ranging from .588 to .941, with a mean descriptive global index of .749. Consequently, no prompt required elimination (Table 4).

Likewise, with regard to inter-rater agreement and reliability, overall agreement among the 34 judges increased substantially between the two rounds. Kendall's coefficient increased from $W = .576$ to $W = .843$ ($p < .001$).

In the second round, agreement was high across all criteria: clarity, $W = .831$; internal coherence, $W = .855$; absence of bias, $W = .882$; linguistic appropriateness, $W = .867$; and precision, $W = .804$ ($p < .001$ in all cases).

Fleiss' κ increased from .432, 95% CI [.391, .465], in the first round to .803, 95% CI [.772, .828], in the second. Ordinal Krippendorff's α , calculated on the same ratings using an ordinal definition of disagreement, showed the same directional pattern, increasing from $\alpha = .561$, 95% CI [.501, .614], to $\alpha = .837$, 95% CI [.808, .861]. The similarity of the pattern obtained using coefficients based on different definitions of agreement provides a sensitivity analysis of the result, without constituting statistically independent corroboration. The complementary analysis using Cohen's κ yielded convergent results. Indeed, across the 561 possible pairs of raters, the median κ increased from .433 (range = .192–.696) to .801 (range = .643–.938). The interpretive scope of these latter results is confined exclusively to the diagnostic assessment of pairwise agreement and not as a substitute for the multi-rater coefficients.

Table 4. Content validity evidence and inter-rater agreement by round

Evidence	Rounda 1	Rounda 2
Aiken's V (range)	.686-.931	.912-.971
Aiken's V (mean)	.855	.939
Lower bound of the 95% CI for V (range)	.591-.865	.841-.917
CVC (range)	.765-.949	.934-.978
CVC (mean)	.891	.954
CVR (range)	.176-.824	.588-.941
Mean CVR	.508	.749
Kendall's W (overall)	.576	.843
Fleiss' κ	.432	.803
Krippendorff's ordinal α	.561	.837
Pairwise Cohen's κ (median)	.433	.801
Pairwise BN (median)	.489	.834
(median) (I)	.923	.978

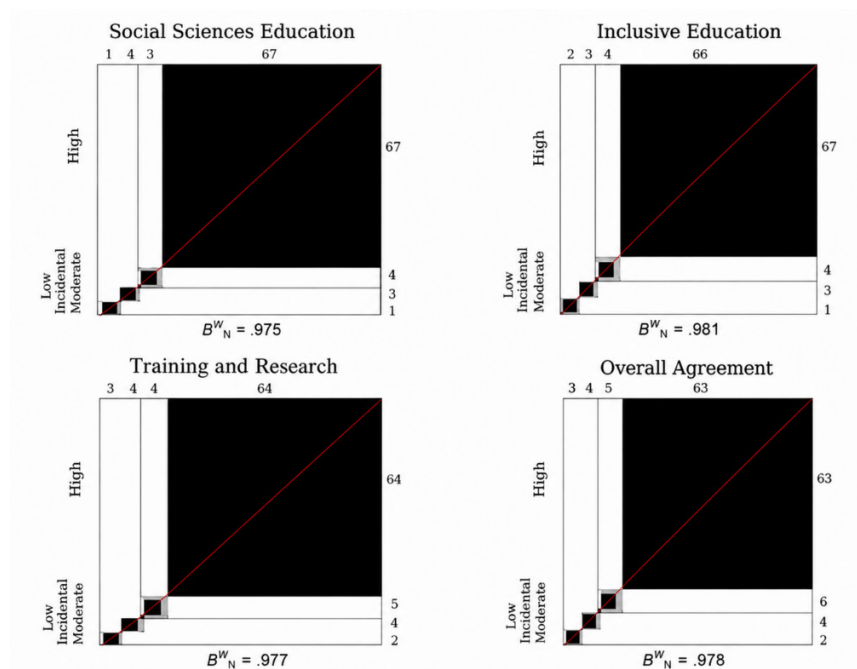
Note. 95% CI = 95% confidence interval; CVC = Content Validity Coefficient; CVR = Content Validity Ratio; κ = kappa; α = Krippendorff's alpha; BN = Bangdiwala's unweighted agreement coefficient; = Bangdiwala's weighted agreement coefficient. In the second round, all 19 prompts were reassessed; four were reformulated and 15 retained their original wording.

Bangdiwala's coefficients likewise showed an improvement in agreement. Considering the 176 within-group pairs, the median BN was .834 (range = .711–.957), and the median weighted B (I) was .978 (range = .956–.995) in the second round. By area of expertise, the median BN / values were .824/.975 in Social Sciences Education, .863/.981 in inclusive education, and .830/.977 in education and research. The greater magnitude of relative to BN should be interpreted in light of the fact that the weighted version assigns partial agreement to discrepancies between adjacent categories on the ordinal scale. For this reason, was not used in isolation to establish the reliability of the instrument.

The Kruskal–Wallis tests identified no statistically significant differences among the three professional groups for any of the five criteria: clarity ($H(2) = 2.410$, $p = .300$, $\epsilon^2 = .013$), coherence ($H(2) = 0.535$, $p = .765$, $\epsilon^2 < .001$), absence of bias ($H(2) = 0.154$, $p = .926$, $\epsilon^2 < .001$), linguistic appropriateness ($H(2) = 0.211$, $p = .900$, $\epsilon^2 < .001$), and precision ($H(2) = 2.385$, $p = .303$, $\epsilon^2 = .012$). Because none of the omnibus tests reached statistical significance, no post hoc comparisons were conducted. These tests refer to the ranks of the individual mean indices derived from the 1–4 coding and should therefore be interpreted as comparisons of the overall rating tendency across areas of expertise, rather than as tests of equivalence between groups.

Taken together, the content indices and agreement coefficients provide complementary perspectives on the expert ratings. Because several of these indicators (Aiken's V, CVC, CVR, Kendall's W, Fleiss' κ , ordinal Krippendorff's α , as well as the complementary pairwise analyses using Cohen's κ and Bangdiwala's coefficients) are derived from the same judgments, their convergence is interpreted in terms of descriptive robustness across different operationalizations rather than as statistically independent evidence. Within this framework, the second version of the PSDI provides consistent evidence of content adequacy and reproducibility of expert ratings.

Figure 2. Within-group Bangdiwala agreement charts and overall agreement



Accordingly, all 19 prompts were retained for the final version of the instrument after the corresponding modifications had been incorporated into the four critical prompts identified during the first round of evaluation.

DISCUSSION

The results support the appropriateness of the procedure adopted to obtain content-based validity evidence for the PSDI and, more specifically, highlight the usefulness of an iterative expert-judgment process guided by a priori decision rules. The first round did not lead to the indiscriminate acceptance of all 19 prompts: four of them (5, 6.1, 10.2, and 11.1) showed insufficient evidence on one or more of the indicators of relevance, representativeness, or essentiality. Following reformulation, all prompts simultaneously met the previously established criteria. Aiken's V increased from a range of .686–.931 ($M = .855$) to .912–.971 ($M = .939$), with lower bounds of the 95% CI ranging from .841 to .917 in the second round; the CVC increased from $M = .891$ to $M = .954$; and the descriptive mean CVR increased from .508 to .749, with final values ranging from .588 to .941. The convergence of these results is methodologically more informative than the isolated interpretation of any single coefficient, as it integrates evidence derived from ordinal ratings of relevance and representativeness, judgments of content essentiality, and technical criteria concerning prompt formulation (Aiken, 1985; Ayre & Scally, 2014; Hernández-Nieto, 2002; Lawshe, 1975; Penfield & Giacobbi, 2004).

The nature of the modifications introduced is consistent with the theoretical domain that the instrument is intended to explore. The neutral reformulation of prompt 5 avoided presupposing a direct and necessary relationship between inclusive education and the development of social thinking, a precaution that is particularly pertinent in light of the absence of causal inferences in the associations and changes related to diversity experiences (Hurtado, 2005; Simons et al., 2023). The greater specificity of prompt 6.1 makes it possible to identify the context, content, methodology, and type of activity, components necessary to distinguish among institutional, teaching, and didactic actions, which the literature characterizes as complementary (Duran & Parra, 2014; Paz Maldonado, 2018). Likewise, reducing the potentially leading nature of prompt 10.2 promotes greater openness in responses concerning faculty positions, whereas disaggregating prompt 11.1 into content, educational experiences, and teaching strategies improves its correspondence with evidence identifying teacher education and reflective methodologies as relevant conditions for addressing diversity educationally (Calderón Jaramillo, 2018; Solbes & Torres Merchán, 2012).

Kendall's W coefficient of agreement increased from $W = .576$ to $W = .843$ and, in the second round, remained high across the five technical criteria ($W = .804$ –.882). Convergenly, Fleiss' κ

increased from .432 (95% CI [.391, .465]) to .803 (95% CI [.772, .828]), whereas ordinal Krippendorff's α , used as the principal estimator given the ordinal nature of the response categories, increased from .561 (95% CI [.501, .614]) to .837 (95% CI [.808, .861]). Pairwise analyses likewise showed a concordant pattern: the median Cohen's κ increased from .433 to .801; Bangdiwala's BN, from .489 to .834; and weighted B, from .923 to .978. The directional pattern obtained with coefficients based on different definitions of agreement shows that the descriptive conclusion is robust to the estimator used. These analyses do not constitute independent verifications, since they were applied to the same set of ratings, nor can any of the coefficients be interpreted in isolation as sufficient evidence of validity (Bangdiwala & Shankar, 2013; Cohen, 1960; Fleiss, 1971; Hayes & Krippendorff, 2007; Krippendorff, 2018; Muñoz & Bangdiwala, 1997). Improvement was likewise observed in the 15 prompts whose wording remained unchanged, although it was descriptively greater in the four reformulated prompts. Because all 19 prompts were reassessed, the design does not allow the specific effect of the modifications to be separated from a possible effect associated with repetition or familiarization. The changes should therefore be interpreted as between-round differences (American Educational Research Association et al., 2014; Hallgren, 2012).

Comparisons across areas of expertise provided complementary evidence based on the absence of statistically significant differences among specialists in Social Sciences Education (social thinking and citizenship), inclusive education, and educational research methods for any of the five technical criteria assessed (clarity, coherence, absence of bias, linguistic appropriateness, and precision), with ε^2 values ranging from $< .001$ to .013. This pattern is compatible with similar technical ratings of the instrument among judges from different disciplinary backgrounds, an aspect of particular relevance for a guide designed to integrate issues related to inclusion, identity, social thinking, and teacher education. Nevertheless, the absence of statistically significant differences should not be interpreted as formal evidence of equivalence between profiles, as the design was not conceived as an equivalence study and the small subgroup sizes limit the precision with which differences of small magnitude can be detected (Kruskal & Wallis, 1952).

The specification matrix shows that the representation of social thinking does not depend on its explicit mention in each prompt, but rather on the cross-cutting elicitation of cognitive-interpersonal, emotional-relational, and critical-analytical components in situations involving inclusion and identity diversity. This organization is consistent with the literature linking social thinking to the analysis of socially relevant problems, perspective-taking, and critical-analytical competencies (Calderón Jaramillo, 2018; Peifer & Taasobshirazi, 2022; Solbes & Torres Merchán, 2012). However, the content-based validity evidence obtained does not allow a factorial structure corresponding to these domains to be inferred, nor does it justify transforming qualitative responses into scores without additional studies specifically designed for that purpose. The PSDI should therefore be understood as a structured guide for qualitative elicitation rather than as a measurement scale.

CONCLUSIONS

The main methodological contribution of the present study lies in the operationalization of a conceptually broad field through prompts subjected to explicit criteria of relevance, essentiality, and technical quality, together with the systematic documentation of the revision process for those prompts that initially failed to meet the established criteria. This procedure promotes greater transparency, comparability, and reproducibility in future interview- and focus-group-based research on identity diversity and inclusion in higher education, a field in which the available studies show considerable heterogeneity both in the procedures employed and in the outcomes examined (López-Ruiz et al., 2021; Nevárez Jiménez et al., 2025).

Likewise, the instrument may be particularly useful for analyzing how faculty members and other university stakeholders establish relationships between inclusion and social thinking, the educational experiences they identify, the feasibility of their practices, and the barriers and priorities they recognize. Nevertheless, its use does not obviate the need to adopt and justify the methodological and analytical decisions inherent to qualitative research, including those concerning sampling, information sufficiency or saturation, researcher reflexivity, category construction, and the consistency of coding procedures, all of which should be specified in each application.

The second version of the PSDI shows a consistent pattern of agreement among expert ratings under different statistical definitions of agreement. Following the reformulation of four prompts, all 19 were reassessed and met the established criteria. Likewise, the overall agreement indicators were higher in the second round, although the design does not allow the observed

change to be attributed exclusively to the textual modifications introduced. These results support considering the revised version a methodologically grounded and sufficiently refined proposal for proceeding to pilot testing in higher education contexts. Its main contribution lies in providing an explicit qualitative structure that has been systematically subjected to expert evaluation for investigating, with greater rigor and systematicity, the relationships among social thinking, identity diversity, inclusive education, and teacher education.

These conclusions should be interpreted in light of several limitations. First, the design provides content-based validity evidence and evidence concerning the reproducibility of expert ratings, but it does not constitute a comprehensive validation of the instrument. To date, the processes through which intended participants understand, interpret, and formulate their responses to the prompts have not been examined, nor have the utility and burden associated with its administration, the stability of its functioning across different contexts, or the potential differences arising from its administration through individual interviews or focus groups. Second, the same 34 judges reassessed all 19 prompts. Although they did not receive aggregate results and completed their ratings independently, repetition prevents the effect of reformulating four prompts from being disentangled from a possible familiarization effect. Consequently, the overall improvement in agreement should not be attributed exclusively to the textual modifications introduced.

Future research should proceed toward pilot testing using cognitive interviews with participants from the intended target populations. It will be necessary to examine the comprehensibility of each prompt, its capacity to generate relevant and nonredundant information, the feasibility of its application across different disciplines and institutions, and, where appropriate, the functional equivalence of linguistically or culturally adapted versions. It would also be advisable to develop a brief administration and coding manual and to examine whether individual interviews and focus groups provide comparable or complementary thematic coverage. If future applications were to derive quantitative scores from the responses, a new instrument-development and validation study specifically designed for that purpose would be essential, in accordance with the Standards for Educational and Psychological Testing (American Educational Research Association et al., 2014).

Referencias

1. Aiken LR. Three coefficients for analyzing the reliability and validity of ratings. *Educational and Psychological Measurement* 1985;45(1):131–142. <https://doi.org/10.1177/0013164485451012>
2. American Educational Research Association, American Psychological Association, National Council on Measurement in Education. *Standards for Educational and Psychological Testing*. American Educational Research Association; 2014. <https://www.testingstandards.net/open-access-files.html>
3. American Educational Research Association. AERA code of ethics: American Educational Research Association approved by the AERA Council February 2011. *Educational Researcher* 2011;40(3):145–156. <https://doi.org/10.3102/0013189X11410403>
4. Ato M, López-García JJ, Benavente A. Un sistema de clasificación de los diseños de investigación en psicología. *Anales de Psicología* 2013;29(3):1038–1059. <https://doi.org/10.6018/analesps.29.3.178511>
5. Ayre C, Scally AJ. Critical values for Lawshe's content validity ratio: Revisiting the original methods of calculation. *Measurement and Evaluation in Counseling and Development* 2014;47(1):79–86. <https://doi.org/10.1177/0748175613513808>
6. Bangdiwala SI, Shankar V. The agreement chart. *BMC Medical Research Methodology* 2013;13:97. <https://doi.org/10.1186/1471-2288-13-97>
7. Betancourth-Zambrano S, Martínez-Daza V, Tabares-Díaz YA. Evaluación de pensamiento crítico en estudiantes de Trabajo Social de la región de Atacama-Chile. *Entramado* 2020;16(1):152–164. <https://doi.org/10.18041/1900-3803/entramado.1.6139>
8. British Educational Research Association. *Ethical Guidelines for Educational Research*. 5th ed. British Educational Research Association; 2024. <https://www.bera.ac.uk/publication/ethical-guidelines-for-educational-research-fifth-edition-2024>
9. Calderón Jaramillo AM. El pensamiento social y los temas sociales relevantes: Una reflexión sobre la formación de los psicólogos. *Nuevas Dimensiones* 2018;7:2–20. <https://doi.org/10.53689/nv.vi7.1>
10. Cohen J. A coefficient of agreement for nominal scales. *Educational and Psychological Measurement* 1960;20(1):37–46. <https://doi.org/10.1177/001316446002000104>
11. Duran S, Parra M. Diversidad cultural para promover el desarrollo de habilidades sociales en educación superior. *Cultura, Educación y Sociedad* 2014;5(1):55–67. <https://revistascientificas.cuc.edu.co/index.php/culturaeducacionysociedad/article/view/995>
12. European Parliament, Council of the European Union. Regulation (EU) 2016/679 of 27 April 2016 on the protection of natural persons with regard to the processing of personal data and on the free movement of such data, and repealing Directive 95/46/EC (General Data Protection Regulation). *Official Journal of the European Union* 2016;L119:1–88. <https://eur-lex.europa.eu/eli/reg/2016/679/oj>
13. Fleiss JL. Measuring nominal scale agreement among many raters. *Psychological Bulletin* 1971;76(5):378–382. <https://doi.org/10.1037/h0031619>
14. Hallgren KA. Computing inter-rater reliability for observational data: An overview and tutorial. *Tutorials in Quantitative Methods for Psychology* 2012;8(1):23–34. <https://doi.org/10.20982/tqmp.08.1.p023>
15. Harpe SE. How to analyze Likert and other rating scale data. *Currents in Pharmacy Teaching and Learning* 2015;7(6):836–850. <https://doi.org/10.1016/j.cptl.2015.08.001>
16. Hayes AF, Krippendorff K. Answering the call for a standard reliability measure for coding data. *Communication Methods and Measures* 2007;1(1):77–89. <https://doi.org/10.1080/19312450709336664>
17. Hernández-Nieto R. *Contributions to Statistical Analysis: The Coefficients of Proportional Variance, Content Validity and Kappa*. BookSurge Publishing; 2002. https://books.google.com/books?id=yk_2PAAACAAJ
18. Hurtado S. Diversity and learning for a pluralistic democracy. In: Allen WR, Bonous-Hammarth M, Teranishi RT, Dano OC, editors. *Higher Education in a Global Society: Achieving Diversity, Equity and Excellence*. *Advances in Education in Diverse Communities*. Vol. 5. Emerald Group Publishing; 2005. p. 249–267. [https://doi.org/10.1016/S1479-358X\(05\)05012-6](https://doi.org/10.1016/S1479-358X(05)05012-6)
19. Kendall MG, Babington Smith B. The problem of m rankings. *The Annals of Mathematical Statistics* 1939;10(3):275–287. <https://doi.org/10.1214/aoms/1177732186>
20. Krippendorff K. *Content Analysis: An Introduction to Its Methodology*. 4th ed. SAGE; 2018. <https://methods.sagepub.com/book/mono/content-analysis-4e/toc>
21. Kruskal WH, Wallis WA. Use of ranks in one-criterion variance analysis. *Journal of the American Statistical Association* 1952;47(260):583–621. <https://doi.org/10.1080/01621459.1952.10483441>
22. Lawshe CH. A quantitative approach to content validity. *Personnel Psychology* 1975;28(4):563–575. <https://doi.org/10.1111/j.1744-6570.1975.tb01393.x>
23. López-Ruiz C, Flores-Flores R, Galindo-Quispe A, Huayta-Franco Y. Pensamiento crítico en estudiantes de educación superior: Una revisión sistemática. *Revista Innova Educación* 2021;3(2):374–385. <https://doi.org/10.35622/j.rie.2021.02.006>
24. Muñoz SR, Bangdiwala SI. Interpretation of kappa and B statistics measures of agreement. *Journal of Applied Statistics* 1997;24(1):105–112. <https://doi.org/10.1080/02664769723918>
25. Nevárez Jiménez LF, Velasco Crespo DR, Martínez Ángulo MA. Impacto de la educación inclusiva en el desarrollo de competencias socioemocionales y su influencia en la convivencia universitaria y el rendimiento académico en los estudiantes universitarios, con un enfoque en la salud mental. *Reincisol* 2025;4(7):3402–3425. [https://doi.org/10.59282/reincisol.V4\(7\)3402-3425](https://doi.org/10.59282/reincisol.V4(7)3402-3425)
26. Ordaz Hernández M. La educación de habilidades sociales desde la Extensión Universitaria: Propuesta de acciones. *Educación en Revista* 2013;29(50):269–283. <https://doi.org/10.1590/S0104-40602013000400017>

27. Ortega-Sánchez D, Bernal López-Sanvicente A. Design, content validity, and inter-observer reliability of the "Digitization of Cultural Heritage, Identities, and Education" (DICHIE) instrument. *Humanities and Social Sciences Communications* 2023;10:53. <https://doi.org/10.1057/s41599-023-01550-z>
28. Paz Maldonado E. Competencias del profesorado universitario para la atención a la diversidad en la educación superior. *Revista Latinoamericana de Educación Inclusiva* 2018;12(2):115–131. <https://doi.org/10.4067/S0718-73782018000200115>
29. Peifer JS, Taasoobshirazi G. Heightened cognitive empathy through ethnic identity and intercultural competence for individuals with historically marginalized identities. *International Journal on Social and Education Sciences* 2022;4(2):290–306. <https://doi.org/10.46328/ijones.324>
30. Penfield RD, Giacobbi PR Jr. Applying a score confidence interval to Aiken's item content-relevance index. *Measurement in Physical Education and Exercise Science* 2004;8(4):213–225. https://doi.org/10.1207/s15327841mpee0804_3
31. Pérez Serrano G, Sarrate Capdevila ML. Diversidad cultural y ciudadanía: Hacia una educación superior inclusiva. *Educación XX1* 2013;16(1):85–104. <https://doi.org/10.5944/EDUCXX1.16.1.718>
32. Simons L, Schoneman S, Hoffman M, Blank N. Equity, inclusion, diversity, and belonging in a multicultural psychology course during the COVID-19 pandemic. In: Mohiuddin M, Aziz MT, Jayashree S, editors. *Multiculturalism and Interculturalism: Managing Diversity in Cross-Cultural Environment*. IntechOpen; 2023. <https://doi.org/10.5772/intechopen.109316>
33. Solbes J, Torres Merchán NY. Análisis de las competencias de pensamiento crítico desde el abordaje de las cuestiones sociocientíficas: Un estudio en el ámbito universitario. *Didáctica de las Ciencias Experimentales y Sociales* 2012;26:247–269. <https://doi.org/10.7203/DCES.26.1928>

Financiación

This research was funded by the project Attitudes, Teaching Disposition, and Interests of Prospective Teachers Toward the Curricular Inclusion of Controversial Issues in Compulsory Secondary Education (ADIFICE) (PID2024-156838OB-I00), funded by the State Research Agency of the Spanish Ministry of Science, Innovation and Universities; by Consolidated Research Unit 285 – Integrated Teaching and Transdisciplinarity (EIT); and by the Recognized Research Group History and Social Sciences Education (DHISO) of the University of Burgos.

Conflictos de interés

The authors declare that they have no conflicts of interest.

Contribución de autoría

Conceptualization: Delfín Ortega-Sánchez; Carlos Pérez-González.

Formal analysis: Delfín Ortega-Sánchez.

Investigation: Delfín Ortega-Sánchez; Carlos Pérez-González.

Methodology: Delfín Ortega-Sánchez.

Validation: Delfín Ortega-Sánchez.

Writing – original draft: Delfín Ortega-Sánchez.

Redacción - revisión y edición: Delfín Ortega-Sánchez; Carlos Pérez-González.