frontiers | Frontiers in Education

TYPE Original Research
PUBLISHED 24 August 2023
DOI 10.3389/feduc.2023.1213358

OPEN ACCESS

EDITED BY
Delfín Ortega-Sánchez,
University of Burgos, Spain

REVIEWED BY
David Parra,
University of Valencia, Spain
Juan Carlos Colomer Rubio,
University of Valencia, Spain
Luis Alves,
University of Porto, Portugal
John Whitehouse,
The University of Melbourne, Australia

*CORRESPONDENCE
Pedro Miralles-Sánchez
✉ pedrmir98@gmail.com

RECEIVED 03 May 2023
ACCEPTED 20 July 2023
PUBLISHED 24 August 2023

CITATION
Miralles-Sánchez P, Gómez-Carrasco CJ and Rodríguez-Medina J (2023) Design and validation of two tools to observe and analyze history lessons in secondary education. *Front. Educ.* 8:1213358.
doi: 10.3389/feduc.2023.1213358

# Design and validation of two tools to observe and analyze history lessons in secondary education

Pedro Miralles-Sánchez[1]*, Cosme J. Gómez-Carrasco[1] and Jairo Rodríguez-Medina[2]

[1] Department of Didactics of Mathematics and Social Sciences, Faculty of Education, University of Murcia, Murcia, Spain, [2] Department of Pedagogy, Faculty of Education, University of Valladolid, Valladolid, Spain

This article focuses on the validation of two data collection instruments, one is the *History Class Observation Tool* (HCOT) and the other is focus groups with students, trainee teachers and tutors from secondary school classrooms. The main objective of the study is to obtain evidence of validity of the two instruments to be used in research on the design, validation, implementation, and evaluation of training units. All this in order to improve the teaching-learning process of historical thinking skills in high school students with the aim of forming critical citizens. The initial set of 32 observable behaviours was reviewed by 9 judges, who rated each on a scale of 0 (strongly disagree) to 3 (strongly agree) in terms of appropriateness, importance, clarity of wording, and observability. As can be seen from the results, both instruments according to the agreement indices suggest that the items, overall, can be considered suitable and important, and observable in the case of the observation instrument, as well as having clarity of wording appropriate to the objectives of both, with high confidence on the part of the experts. If we compare it with similar studies that we have discussed previously, we can see how this validation process has been quite rigorous and novel following some guidelines set previously in certain studies.

KEYWORDS

**history, secondary education, baccalaureate, validation, systematic observation, historical thinking, training units**

## Introduction

This article is part of the research "The teaching and learning of historical competences in baccalaureate: a challenge to achieve a critical and democratic citizenship" based on the design, validation, implementation, and evaluation of training units to improve the teaching-learning process of historical thinking skills in baccalaureate students aimed at the formation of critical citizens. The aim of this article is therefore to obtain evidence of the content validity of two data collection instruments, namely, the *History Class Observation Tool* (HCOT) and the focus groups with students, trainee teachers, and tutors in secondary school classrooms.

The development and implementation of observation instruments can be very useful to effectively design training programmes and evaluate classroom interventions. However, most of these instruments focus on teachers' generic competences rather than subject-specific competences. Therefore, some researchers have highlighted the importance of designing specific observation instruments in research on teacher education and competences (Desimone, 2009; Schoenfeld, 2013). However, to date, there are no validated and reliable observation instruments for analyzing teaching-learning processes in history. This is unfortunate, especially because, as Sáiz Serrano and Gómez Carrasco (2016) and Van Boxtel et al. (2020) warn, current teacher education programmes may not meet the needs of history teachers to achieve the objectives set out in the curricula. Observational instruments that assess the teaching strategies of history teachers could allow the identification of specific needs and thus facilitate the design of teacher education plans and/or programmes, which is an important and novel contribution to the field of history teacher education.

As Huijgen et al. (2017) point out, the use of standardized observation instruments in history education research is an under-addressed topic, and, in particular, instruments for observing strategies for developing historical thinking in the classroom are not available. Since the 1970s, increasing attention to the assessment of teachers' generic competencies has led to the development of a variety of observation instruments that are widely used to assess primary and secondary education, such as the *Stallings Observe System* (Stallings and Kaskowitz, 1974), the *Framework for Teaching* (Danielson, 1996), the *International System for Teacher Observation and Feedback* (Teddlie et al., 2006), the *International Comparative Analysis of Learning and Teaching* (Van de Grift, 2007), and the *Classroom Assessment Scoring System* (Pianta et al., 2008, 2011), among others. Although some recently developed observation instruments focus on more specific teaching competences, such as classroom conversation (Mercer, 2010), project-based learning (Stearns et al., 2012), and learning and instructional reform (Sawada et al., 2002), only a few observation instruments focus on teachers' strategies in specific subjects, such as reading in English (Gertsen et al., 2005), content and language integrated learning (De Graaff et al., 2007), English language arts (Grossman et al., 2010), and mathematics teaching (Matsumura et al., 2008; Hill et al., 2012; Schoenfeld, 2013).

In terms of observation instruments used in history, we can first highlight the pioneering work of Nokes (2010), which focused on history teachers' literacy-related decisions about the types of texts they used and how students were taught to learn with these texts. In this study, two observation instruments were created: one to record the type of texts teachers used and one to record the activities and instruction they provided. To create the text log sheet, a group of experienced secondary school history teachers generated a list of common types of resources they might use in class, listing each as a row. To create the activity record sheet, the same group followed the same procedure with a list of common activities in history classrooms. Both forms provided space for adding texts or unplanned teaching activities. On both recording sheets, the 90-min class session was divided into six columns representing 15-min time units. Detailed instructions for the use of the recording sheets were drawn up, along with a description of what could and could not be ticked in a certain box. Moreover, it was analyzed in four phases. First, the frequency of use of various texts and didactic activities was calculated. In the second phase of the analysis, differences between teachers were investigated. Third, an analysis was carried out to see how each teacher used each type of text. Fourth, based on the frequency counts, teachers were placed on a spectrum showing the proportion of instruction on historical narrative and the amount of instructional time on historical processes (Nokes, 2010).

But a key observational instrument more closely related to historical thinking was created by Van Hover et al. (2012) and called the *Protocol for Assessing the Teaching of History* (PATH). This instrument provides a lens through which to observe secondary history teaching with the aim of providing a means for structured and focused observation with the goal of improving instruction, although it was not intended to provide guidelines on how to teach and learn history. PATH initiates the conversation about how to capture and explore specific teaching behaviors. In terms of validation, history educators (in the United States and the United Kingdom) and measurement experts reviewed the dimensions and provided critical comments and suggestions. At the same time, the authors watched hundreds of hours of videos of history teaching in secondary schools (Van Hover et al., 2012).

It is based on the Classroom Assessment Scoring System-Secondary (CLASS-S) (Pianta et al., 2008, 2011), an instrument developed to assess classroom quality. CLASS-S focuses on student-teacher interactions as the primary mechanism for student learning, and PATH uses the same structure and scoring/coding approach. Prior to using the tool, PATH coders are trained on each dimension of a rubric through a detailed manual that describes the specific teaching behaviors that make up each dimension. The high inference instrument is scored on a 7-point rating scale based on alignment with the anchor descriptions at Low (1, 2), Medium (3, 4, 5), and High (6, 7). In addition, to develop the discipline-specific dimensions, they first conducted an extensive review of the literature on history teaching, looking for work that could help identify observable teacher and student behaviors that contribute to student learning (Van Hover et al., 2012).

Six separate dimensions emerged from this literature review. Lesson components: Assesses the structure and flow of the history lesson, paying attention to objectives, assessment, and appropriate instructional approaches. In addition, it assesses attention to an overarching concept or framing a historical question. Comprehension: Assesses whether students understand the framework, key concepts, and content of the story and whether they are able to express this knowledge in different ways. Narrative: Assesses the structure and flow of a narrative and whether students understand chronology, context, cause and effect, and how narratives are constructed. Narrative is defined as any contemporary verbal or written account (could include texts, lectures, websites, or films). Interpretation: Assesses the level of attention paid to the fluid and controversial nature of the story, as well as consideration of (if appropriate to the lesson objectives) agency, meaning, diverse points of view, and recognition of perspective. Sources: Assesses the selection, accessibility, purpose, and level of analysis of historical sources used in the classroom and whether there is an opportunity for meaningful historical research. Historical practices: Assesses whether general instructional practices (writing, discussion, and simulations) are implemented in ways that are authentic and appropriate to the discipline (Van Hover et al., 2012).

Gestsdóttir et al. (2018) underline the fact that the PATH is still under development and, despite the importance of the definition of the six dimensions, none of them is adequate for providing an overview of teacher behavior that reinforces students' historical thinking and reasoning. Therefore, there is a clear need for a more comprehensive instrument that continues to focus on the specific components of history teaching. They developed and evaluated the Teach-HTR (Historical Thinking and Reasoning) observation instrument in four phases: (1) literature review,

(2) expert consultation, (3) first pilot of the instrument, and (4) second pilot of the instrument. This instrument examines lessons with high and low scores to explore the potential of the instrument to give teachers feedback on what they are already doing and where there is room for development (Gestsdóttir et al., 2018).

They define seven categories: *Communicating goals related to historical thinking and reasoning*; *Demonstrating historical thinking and reasoning*; *Using sources to support historical thinking and reasoning*; *Presenting multiple perspectives and interpretations*; *Explicit instructions on historical thinking and reasoning strategies*; *Engaging students in individual or group tasks that require historical thinking and reasoning*; and *Engaging students in a whole-class discussion that asks for historical thinking and reasoning* (Gestsdóttir et al., 2018).

Another key instrument is that of Huijgen et al. (2017), from which we have taken some categories for the observation instrument we have taken. They developed and tested a domain-specific observation instrument focusing on historical contextualization called the *Framework for Analyzing the Teaching of Historical Contextualization* (FAT-HC). Their instrument was based on four teaching strategies for historical contextualization. The first strategy is the reconstruction of the historical context. Students must have knowledge of the historical context, including knowledge of chronology and space, and of socio-economic, socio-cultural, and socio-political developments before they can successfully carry out historical contextualization. The second strategy is to enhance historical empathy, e.g., by selecting a historical agent relevant to the topic under study, focusing on the role and position of the historical agent in society, and promoting students' affective connections with the historical agent. The third strategy is to enhance the use of knowledge of the historical context. Students not only have to reconstruct a historical context but should also use it, for example, to determine causes and consequences, compare historical phenomena, and understand different perspectives on phenomena. The last strategy is to enhance the awareness of present-oriented perspectives among pupils when they examine the past. Without awareness of the differences between the past and the present, students are not able to compare, explain, or evaluate the past (Huijgen et al., 2019).

They modeled their instrument on Van de Grift (2007, 2009) *International Comparative Analysis of Learning and Teaching* (ICALT) observation instrument, resulting in a total list of 45 items in the first version of the FAT-HC. The aim of the study was to develop a reliable observation instrument and a scoring design to assess how history teachers promote historical contextualization in classrooms. Using expert panels, they found positive indicators of the content validity of the instrument, and by analyzing generality theory, they found indicators that the instrument is unidimensional, as it showed that a large proportion of the variance of the instrument was explained by differences between observed teachers and a small proportion of the variance was explained by differences in lessons and observers. They also organized two expert roundtables to ensure the face and content validity of the instrument. Finally, they trained 10 history students in the use of the observation instrument and observed a videotaped history lesson using the instrument, calculating Cronbach's alpha for their observation scores to explore the internal consistency of the instrument (Huijgen et al., 2017).

Finally, it is worth noting the more recent work of Oattes et al. (2022), who used three instruments to collect data. First, a quantitative *Pedagogical Content Knowledge* (PCK) checklist was used to record the frequency and quality with which particular PCK items were used, supplemented by qualitative data software to analyze fourteen key items from twelve paired lessons to distinguish differences and similarities in the language of instruction. They highlight that, for a quantitative analysis of history teachers' application of PCK, existing models of observation of teaching behavior are general education-oriented and appeared to be either too pedagogical-didactic in general (Van de Grift, 2007), too language-oriented (De Graaff et al., 2007), or too intellectually demanding for the younger learners involved. They concretised them to analyze the classroom teaching of history teachers using PCK using Monte-Sano and Budano's (2013) *Framing History*. Finally, they used the Protocol for the Assessment of Teaching History (PATH), designed by Van Hover et al. (2012), with the six categories it includes with the aim of improving instruction. In addition, for the quantitative part, the assessment scores of the 24 observed classes were analyzed using the SPSS software to calculate descriptive statistics, quantifying the differences between the applications of the PCK categories (Oattes et al., 2022).

In contrast to the instruments outlined earlier, the observation instrument we present here focuses on a unique but very important competence for history teachers, namely the fostering of historical thinking skills, which are embedded in history curricula worldwide (Van Drie and van Boxtel, 2008; Seixas and Morton, 2013). In previous studies, we have analyzed the impact of a training programme in the Geography and History specialization of the Master's Degree in Teacher Education on the motivation, satisfaction, and perception of learning of history students (Gómez Carrasco et al., 2020, 2021), and we have analyzed the teaching approaches of history teachers in Spain and their relationship with their views on the use of digital resources in a classroom (Gómez Carrasco et al., 2022). The data obtained through the observation instrument that we will design will allow us to complement these previous works, based on self-reported measures, with a micro-analytical perspective that provides greater richness and detail of what really happens in the classroom. Moreover, the combination of both techniques (systematic observation/self-report) will allow us to analyze the relationship between teachers' beliefs and their practices for the development of historical thinking skills, opening up a promising line of research as suggested by Huijgen et al. (2019).

## Objective

The aim of this article is to obtain evidence of the content validity of two data collection instruments, namely, the *History Class Observation Tool* (HCOT) and the focus groups with students, trainee teachers, and tutors in secondary school classrooms. These instruments will be used in the research "The teaching and learning of historical competences in baccalaureate: a challenge to achieve a critical and democratic citizenship" based on the design, validation, implementation, and evaluation of training units to improve the teaching-learning process of historical thinking competences in baccalaureate students aimed at the formation of critical citizens. It is evaluative research with a mixed explanatory approach, a quasi-experimental design with an experimental group and a control group, and the use of quantitative and qualitative methods and observation.

# Methods

## Research design

The initial set of 32 observable behaviors was reviewed by nine judges, who rated each on a scale of 0 (strongly disagree) to 3 (strongly agree) in terms of appropriateness, importance, clarity of wording, and observability. Similarly, the 41 questions posed for the focus groups were rated in terms of appropriateness, importance, and clarity of wording on a scale of agreement between 0 (do not agree at all) and 3 (strongly agree) by 8 expert judges. To analyze the agreement among the judges, Bangdiwala's weighted coefficients of agreement ($B_{\mathrm{N}}^{\mathrm{W}}$) (Bangdiwala, 1987) were calculated. Bangdiwala's $B_{\mathrm{N}}^{\mathrm{W}}$ agreement index allows a graphical representation of the degree of agreement and provides a measure of the strength of agreement. In this representation, the black squares show observed agreement, while the gray areas represent partial agreement. The white area of each rectangle is the graphical representation of disagreement. Data were analyzed using the R software v. 4.0.4 (R Core Team, 2021).

## Instruments

First, the *History Class Observation Tool* (HCOT) is composed of three dimensions, namely, Teaching Discourse (Verbal), Teaching Materials, and Student Activity. The first dimension, Teaching Discourse, is subdivided into five categories, namely, Exploration and Activation of Prior Knowledge; Contextualization; Interpretation; Historical Thinking; and Teaching Methods, Strategies, and Techniques. This gives a total of 38 items (see Annexes).

The focus groups consisted of interviews with students, trainee teachers, and secondary school tutors, with a total of 16, 9, and 16 questions, respectively (see Annexes).

These instruments will be used in the four phases of the project: pre-observation of the classroom (I), design of training units (II), implementation of training units (III), and evaluation of results (IV). Validation of these instruments would be essential to ensure that the data collected are accurate and reliable. One way to validate the instruments would be through review by experts in the field and pilot testing on a small group of participants to assess the effectiveness and relevance of the questions and observation procedures.

# Results

## Inter-judge agreement systematic history classroom observation instrument and focus groups

Table 1 shows Bangdiwala's strict ($\mathrm{B}_N$) and weighted or partial agreement ($\mathrm{B}^W{}_N$) coefficients (Bangdiwala, 1987; Friendly and Meyer, 2016) obtained for both the observation instrument and the focus groups. The BN coefficients (on the values of the matching matrix that is subjected to the concordance analysis) are calculated using the formula:

$$B_N = \frac{\sum \mathrm{n}_{\mathrm{ii}}^2}{\sum \mathrm{n}_{\mathrm{i+}}\mathrm{n}_{\mathrm{+i}}} = B_N$$

$$= \frac{\text{\textit{área de los cuadrados negros}}}{\text{\textit{área total de los cuadrados de cada categoría}}}$$

To account for partial agreement ($\mathrm{B}^W{}_N$), since this is an ordinal rating scale, a weighted contribution of the off-diagonal cells is included as a function of the steps (separation) from the main diagonal. These partial agreements are included in the graph as squares of a lighter shade (gray squares) than the strict agreement (black squares). So a pattern of weights (weights w1, w2, ..., wb), according to the formula proposed by Fleiss and Cohen (1973), is applied to the shaded areas separated by $b$ steps from the diagonal. Thus, the following formula is used to calculate the partial agreement coefficient ($\mathrm{B}^W{}_N$):

$$B_N = 1 - \frac{\sum[\mathrm{n}_{\mathrm{i+}}\mathrm{n}_{\mathrm{+i}} - \mathrm{n}_{\mathrm{ii}}^2 - \sum w_b A_{bi}]}{\sum \mathrm{n}_{\mathrm{i+}}\mathrm{n}_{\mathrm{+i}}}$$

Where $w_b$ represents the weighting as a function of distance from the diagonal and $A_{bi}$ represents the area of the shaded areas with a separation $b$ from the diagonal. As shown in Figure 1, the weighted Bangdiwala concordance coefficients (Bangdiwala, 1987) obtained were $\mathrm{B}^W{}_N = 0.94$ (appropriateness); $\mathrm{B}^W{}_N = 0.94$ (importance); $\mathrm{B}^W{}_N = 0.92$ (clarity of wording); and $\mathrm{B}^W{}_N = 0.88$ (observability). To interpret the agreement coefficients, Muñoz and Bangdiwala (1997) propose the following criteria: values below 0.09 indicate a poor level of agreement; between 0.09 and 0.25, poor agreement; between 0.25 and 0.49, moderate agreement; between 0.49 and 0.801, good agreement; and above 0.801, excellent agreement. The results obtained, according to this interpretation, suggest an excellent level of agreement, so that the items, overall, can be considered adequate, important, and observable. Moreover, the clarity of the wording, in the opinion of the experts, is adequate for observing the effectiveness of a formative programme for teaching history at the baccalaureate level based on epistemological and methodological changes.

To complete the analysis, we also calculated the coefficients of agreement AC2 (Table 1) proposed by Gwet (2008) to overcome the limitations and paradoxes of Cohen's Kappa coefficient (Cicchetti and Feinstein, 1990; Feinstien and Cicchetti, 1990), especially in situations where the degree of agreement is high. In this case, AC2 was used, as it allows for partial agreement in the case of ordinal data. The AC2 agreement indices obtained can be considered excellent for all the variables examined according to the criteria proposed by Muñoz and Bangdiwala (1997), or taking into consideration the proposal by Gwet (2021) to interpret the values obtained by calculating the probability of belonging to each of the intervals (Interval Membership Probability), the results of which are shown in Tables 2, 3, the cumulative probabilities for the AC2 agreement coefficient using the reference scale proposed by Muñoz and Bangdiwala (1997). Based on these results, the level of

TABLE 1 Bangdiwala's stringent ($B_N$) and weighted ($B_N^W$) agreement coefficients.

| | Coefficient | Suitability | Importance | Clarity | Observability |
|---|---|---|---|---|---|
| Observation instrument | $B_N$ | 0.835 | 0.801 | 0.856 | 0.736 |
| | $B^W{}_N$ | 0.942 | 0.944 | 0.918 | 0.880 |
| | $AC_2$ | 0.950 | 0.946 | 0.943 | 0.843 |
| Focus groups | $B_N$ | 0.798 | 0.777 | 0.646 | – |
| | $B^W{}_N$ | 0.907 | 0.924 | 0.838 | – |
| | $AC_2$ | 0.881 | 0.897 | 0.849 | – |

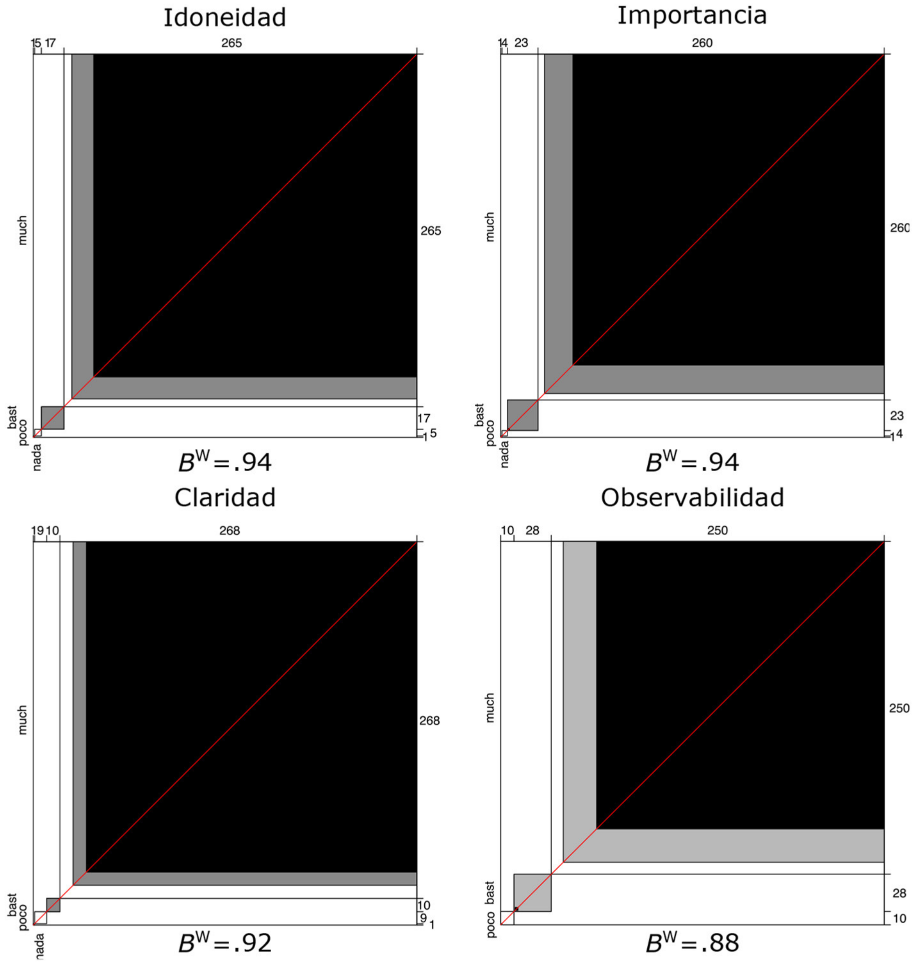

FIGURE 1
Results of the Bangdiwala judges agreement analysis—observation instrument.

agreement obtained for the behaviors covered by the observation instrument can be considered excellent with a confidence of over 95% for all the variables analyzed. As far as the focus groups are concerned, agreement can be considered excellent for the appropriateness and relevance of the items asked, with more than 95% confidence, and good or better with 100% confidence for the clarity of the wording of the items.

Next, the responses of the same set of judges on their assessment of the questions posed to the focus groups were analyzed. In this case, as can be seen in Figure 2, the weighted Bangdiwala agreement coefficients (Bangdiwala, 1987) obtained were: $B_N^W = 0.91$ (appropriateness); $B_N^W = 0.92$ (importance); $B_N^W = 0.84$ (clarity of wording). These indices of agreement suggest that the items, overall, can be considered suitable and important, and also that the clarity of the wording, in the opinion of the experts, is adequate to pose the questions proposed in the focus groups with the aim of identifying changes and permanence in teaching practices, in the role of the students, and in the learning of historical competences within the group, and to identify the role of the school

TABLE 2 Cumulative probabilities of membership in benchmark ranges—observation instrument.

| Agreement values | Interpretation | Suitability | Importance | Clarity | Observability |
|---|---|---|---|---|---|
| | | 0.950<br>(SE = 0.011) | 0.946<br>(SE = 0.012) | 0.943<br>(SE = 0.012) | 0.842<br>(SE = 0.022) |
| | | 95% CI<br>(0.927–0.974) | 95% CI<br>(0.922–0.970) | 95% CI<br>(0.919–0.969) | 95% CI<br>(0.786–0.900) |
| (0.81–1) | Excellent | 0.999 | 0.999 | 0.999 | 0.970 |
| (0.49–0.81) | Good | 1 | 1 | 1 | 1 |
| (0.25–0.49) | Moderate | 1 | 1 | 1 | 1 |
| (0.09–0.25) | Scarce | 1 | 1 | 1 | 1 |
| (0.01–0.09) | Poor | 1 | 1 | 1 | 1 |

TABLE 3 Cumulative probabilities of membership in benchmark ranges—focus groups.

| Agreement values | Interpretation | Suitability | Importance | Clarity |
|---|---|---|---|---|
| | | 0.881<br>(SE = 0.024) | 0.897<br>(SE = 0.018) | 0.849<br>(SE = 0.026) |
| | | 95% CI (0.832–0.930) | 95% CI (0.860–0.934) | 95% CI (0.797–0.903) |
| (0.81–1) | Excellent | 0.998 | 0.999 | 0.933 |
| (0.49–0.81) | Good | 0.999 | 1 | 1 |
| (0.25–0.49) | Moderate | 1 | 1 | 1 |
| (0.09–0.25) | Scarce | 1 | 1 | 1 |
| (0.01–0.09) | Poor | 1 | 1 | 1 |

context, the students, and the teaching staff in the results of the experimentation.

Subsequently, the *Content Validity Ratio* (CVR) (Lawshe, 1975) was calculated for each of the behaviors included in the observation instrument and for each of the focus group questions, for each of the variables analyzed (appropriateness, importance, clarity of wording, and observability), and the *Content Validity Index* (CVI) (Lawshe, 1975) for the set of items.

This is an indicator of inter-judge agreement that can take values between −1 (total disagreement) and +1 (total agreement), so that the CVR value is negative if agreement occurs with less than half of the judges; CVR is zero if there is exactly half agreement among the expert judges; and CVR is positive if more than half of the judges agree on the item rating. For the interpretation of the results with nine judges, Ayre and Scally (2014) propose the critical value of CVR = 0.778 ($p = 0.020$), which assumes that at least eight of the nine judges agree on the item rating and exceeds the probability of agreement by chance effect at a 95% confidence level ($\alpha = 0.05$), while for eight judges, the critical value of CVR proposed by these authors is 0.750 ($p = 0.035$).

For the calculation of the CVR values, the ratings of the judges who selected options 2 (quite a lot) and 3 (a lot) were grouped together, and the ratios were calculated using the formula proposed by Lawshe (1975):

$$RVC = \frac{n_e - (\frac{N}{2})}{(\frac{N}{2})}$$

where $n_e$ is equal to the number of judges who consider the item to be quite or very adequate, important, clear, or observable, and $N$ is the total number of judges. Finally, the Content Validity Index (CVI) was obtained by calculating the average content validity (CVR) of each of the items in each of the variables considered globally.

With regard to the behaviors present in the observation instrument, the RVC values ranged from 0.78 to 1 for the four variables considered (appropriateness, importance, clarity of wording, and observability), so that all exceeded the critical RVC value proposed by Ayre and Scally (2014) of 0.778 for nine judges. The overall CVI for the set of behaviors was 0.96 (appropriateness), 0.97 (importance), 0.93 (clarity), and 0.93 (observability).

Regarding the questions posed for the focus groups, the RVC values for all items exceeded the critical value of 0.750 ($p = 0.035$), except for item 14 of the focus group of active teachers in suitability (RVC = 0.5) and clarity (RVC = 0.5) (14. Why do you think that these practices are not reproduced on a daily basis in the classroom? They are not useful/lack of training/school culture/lack of time associated with excessive content, class hours, ratios, bureaucracy, time to prepare classes...?); item 16 of the focus group for students in suitability (RVC = 0.5) (16.5) (16. Do you think that this way of working with history helps us to be better citizens?); item 6 of the focus group of in-service teachers on clarity (RVC 0.5) (6. Do you consider that this methodology brings about changes in the richness of student learning? What kind of changes?); and item 8 of the focus group of in-service teachers on clarity (RVC 0.25) (8. In what way could you have acted to achieve better results in order to achieve better results in the development of historical competences by the

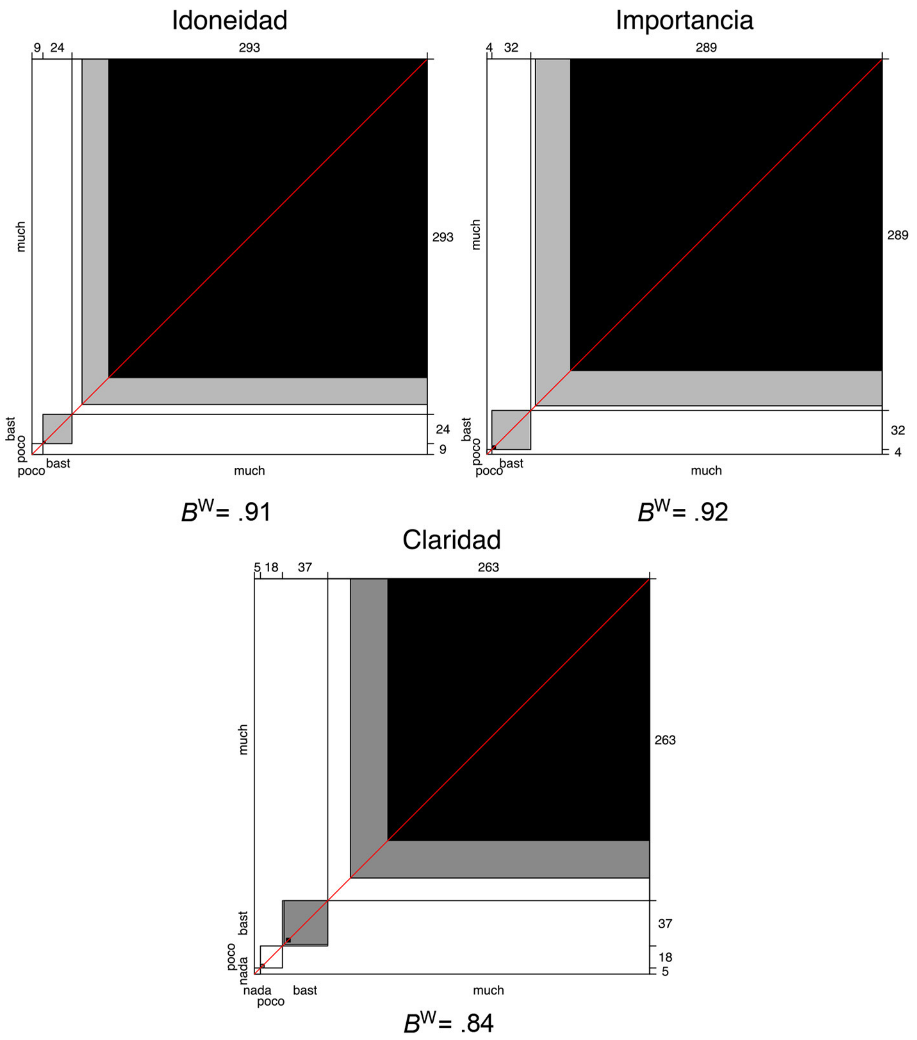

FIGURE 2
Results of the Bangdiwala judges agreement analysis—focus groups.

students?). The CVI values for the set of items proposed for the focus groups were 0.93 (appropriateness), 0.96 (importance) and 0.84 (clarity).

In light of these results, the judges' qualitative assessments of the items noted that did not pass the content validity ratio threshold were reviewed. With regard to the clarity of the wording of item 8 of the in-service teacher focus group, an error in the wording was identified and corrected ("to achieve better results to achieve better results"). Four of the eight judges identified this error and pointed it out in their comments.

In relation to item 6 of the focus group of active teachers in clarity (6. Do you think that this methodology brings changes in the richness of student learning? What kind of changes?), the experts pointed out that it would be convenient to explain what is considered "richness" (e.g., j4—"I find the expression 'richness of learning' abstract and somewhat confusing"; j1—"What is considered as richness?"). Regarding item 16 of the student focus group (16. Do you think that this way of working with history helps you to be a better citizen?), the judges expressed the possibility that students may have difficulties in understanding the meaning of the question (e.g., j1—"In what sense?"; j8—it would be convenient to "Ask about what they consider to be 'good' or 'better' citizens". j2—"Formulation somewhat misleading because of the comparison between better and worse"; j5—"I would speak of 'citizens' and include at the end the question 'Why?' I consider it essential that they explain the causes in order to check what they understand by citizenship and what aspects they focus on, as there may be wide divergences").

Why do you think that these practices are not reproduced on a day-to-day basis in the classroom? They are not useful/lack of training/school culture/lack of time associated with excessive content, class hours, ratios, bureaucracy, time to prepare classes ...? The experts considered that it would be appropriate to ask two different questions (e.g., j1—"Question too broad. It would be convenient to divide it in order to cover everything in the answers; otherwise, it is possible that some aspects are left out"; j3—"I would ask two different questions") or that it is a "biased" question (e.g., j7—"It is a biased question: it conditions that they see that they do not apply, there will be some who apply some strategies in this respect, or at least that they identify themselves and have had experiences in this respect"). Therefore, it seems appropriate to revise the wording.

## Discussion and conclusion

As can be seen from the results, both instruments, according to the agreement indices, suggest that the items, overall, can be considered suitable, important, and observable in the case of the observation instrument, as well as having clarity of wording appropriate to the objectives of both, with high confidence on the part of the experts. Regarding the behaviors present in the observation instrument, the RVC values ranged between 0.78 and 1 for the four variables considered (appropriateness, importance, clarity of wording, and observability). About the questions posed for the focus groups, the RVC values for all the items exceeded the critical value of 0.750 ($p = 0.035$), except for those indicated above, which would be the lines of improvement of the research.

If we compare it with similar studies that we have discussed previously, we can see how this validation process has been quite rigorous and novel, following some guidelines set previously in certain studies. It should be noted that, due to the specificity of history education, the use of observation instruments has been infrequent and underestimated (Van Hover et al., 2012). However, this trend is beginning to change, as we can see an evolution from a purely qualitative observational approach to a mixed approach such as ours, although the evolution is not chronological.

Similarities can be found in work such as that of van Hover, Hicks, and Cotton (PATH), which used history educators and measurement experts to review the dimensions, providing critical comments and suggestions, while also using the resource of viewing hundreds of hours of videos of history teaching in secondary schools (Van Hover et al., 2012). Also noteworthy is the Teach-HTR observation instrument by Gestsdóttir et al. (2018), who reviewed literature, consulted experts, and conducted two pilots of the instrument for validation. Huijgen et al.'s instrument in the FAT-HC, a major reference for the development of our observation instrument, took this a step further by using expert panels to find positive indicators of the content validity of the instrument, as well as analyzing the theory of generality to have indicators that the instrument is unidimensional. They also organized two expert roundtables to ensure the face and content validity of the instrument, and, finally, they trained 10 History students in the use of the observation instrument and observed a videotaped History class using it, calculating Cronbach's alpha for their observation scores in order to explore the internal consistency of the instrument (Huijgen et al., 2017).

The final objective is to design, validate, implement, and evaluate the effectiveness of training units to improve the teaching-learning process of historical thinking skills in Baccalaureate students in order to train critical citizens. It should be remembered that historical thinking is a didactic approach that aims to teach students to *think historically* by deploying different strategies and skills to analyze and respond to different historical questions and to understand the past in a more complex way. To learn about history, we must resort to the use of skills focused on reflection, analysis, argumentation, and interpretation of the past. Such skills are not innate; therefore, they must be acquired and developed in the classroom (Chapman, 2011; Gómez Carrasco et al., 2014). Seixas and Morton (2013) state that historical thinking is a creative process developed by historians to generate new historical narratives through the analysis of sources from the past. These competences and strategies are related to the search for, selection, and treatment of historical sources, empathy, multi-causal explanation, and historical perspective (Peck and Seixas, 2008; Seixas and Morton, 2013).

The importance of teaching historical thinking in the classroom lies in the fact that historical thinking does not develop naturally but needs explicit teaching (Wineburg, 2001). The central core of this theoretical approach is occupied by a small group of methodological concepts that identify the historian's own ways of working. These concepts are variable and do not form a closed and invariable list, but each author attaches greater importance to certain aspects. Some of the historian's most characteristic ways of working include the use of sources and evidence, changes and continuities, empathy and historical contextualization, causes and consequences, and narratives and interpretations. These concepts of historical thinking play a transcendental role in the assessment framework of historical competences (Santiesteban Fernández, 2010; Gómez Carrasco et al., 2017).

Understanding history involves understanding these categories and processes of historical thought. The assessment model for Geography and History should encourage students to reflect on historical content. It is necessary to establish a cognitive model for learning history in order to correctly assess historical knowledge (Carretero and López, 2009; Carretero and Van Alphen, 2014). This cognitive model that we are going to develop must have appropriate techniques and instruments for assessing first- and second-order historical content and skills (Domínguez Castillo, 2015). This requires the collaboration of various social and human disciplines, such as history, art, geography, and literature.

In terms of identifying teaching models, it is worth highlighting the line of research developed by Trigwell and Prosser (2004) based on interviews with teachers and a questionnaire called the Approaches to Teaching Inventory (ATI) (Trigwell et al., 2005). They identified four different conceptions of teaching and three methodologies, establishing five approaches that can be grouped into three broad models or ways of teaching. In the first model, the role of the teacher is greater, since the importance lies in the transmission of content, students assume a passive role, limiting themselves to receiving and memorizing the knowledge transmitted by teachers, thus establishing a unidirectional relationship without considering their experience, previous knowledge, characteristics, or context. The most commonly used methodological strategy is the master class, and the main resources used are the textbook and class notes. In addition, a final examination of the learning contents is usually established (Galvis, 2007; Castejón et al., 2009; Hernández et al., 2012; Guerrero-Romera et al., 2022).

On the other hand, there is learner-centered teaching, which differs from the previous one in that the teacher's intention is to provoke conceptual change and intellectual growth in the learner. Thus, the teacher acts as a guide, guiding students in the process of constructing their own knowledge, encouraging their conceptions, and providing them with opportunities to interact, debate, investigate, and reflect. The aim of this model is for students to learn content by questioning and reflecting on it. The strategies employed are active and inquiry-based. In contrast to the previous model, which encouraged competitiveness and individualism, this approach favors interaction and cooperation between the individuals involved in the teaching and learning process and prioritizes continuous assessment (Vermunt and Verloop, 1999; Kember and Kwan, 2000; Trigwell et al., 2005; Henze and van Driel, 2011). Finally, there is a third, intermediate model based on teacher-student interaction, although it should be noted that there is a hierarchical relationship between the different approaches, with each including elements of the previous one (Guerrero-Romera et al., 2022).

To conclude, this proposal represents a significant improvement compared to the traditional methods used in social science teaching, as it seeks to develop essential skills for critical thinking and citizenship training, and its effectiveness is also evaluated through rigorous methods and a scientific approach. To develop the competences associated with historical thinking, the introduction of the historian's method and techniques and historical awareness are key elements (Domínguez Castillo, 2015). This requires a methodological change in the approach to social sciences classes with the use of a greater variety of techniques beyond the mere expository master class. All of this is to encourage a critical spirit and autonomous learning, and therefore the formation of critical and independent citizens who know how to judge for themselves the vicissitudes that civic life in a democracy demands of them.

There is still an overuse of textbooks and the expository strategy by teachers who teach history (Sobejano and Torres, 2009; Valls and López, 2011; López and Valls, 2012; Carretero and Van Alphen, 2014; Colomer et al., 2018). However, more and more teachers in Spain are in favor of a teaching model in which the student acquires a greater role through the implementation of innovative resources (heritage, written and oral sources, new technologies) and educational strategies that encourage the active participation of students in the teaching and learning process (project-based learning, gamification, and flipped classroom) (Olmos, 2017; Gómez et al., 2018; Gómez Carrasco et al., 2020; Sánchez et al., 2020). Therefore, it is important to be aware of developments in the incorporation of competency-based social sciences teaching and a learner-centered model at all levels of education. For this reason, it is necessary to analyze the teaching profiles of history, geography, and art history teachers by means of observation instruments that make it possible to describe and analyze their classroom practices (Guerrero-Romera et al., 2022).

## Data availability statement

The original contributions presented in the study are included in the article/Supplementary material, further inquiries can be directed to the corresponding authors.

## Ethics statement

Ethical issues were carefully contemplated in this study. Participants were informed about the objectives and procedures of the study and how their rights were going to be protected. Participation in the research was voluntary and anonymous.

## Author contributions

CG-C and JR-M conceived and designed the project and doctoral thesis of which this study is part, have made methodology, validation, data collection, and formal analysis. PM-S and JR-M have co-written the manuscript and contributed to revisions, having read and approved the submitted manuscript. All authors have read and agreed to the published version of the manuscript. All authors contributed to the article and approved the submitted version.

## Funding

This publication is part of grant PRE2021-097619, funded by MCIN/AEI/10.13039/501100011033 and ESF+. It is part of the research project La enseñanza y el aprendizaje de competencias históricas en bachillerato: un reto para lograr una ciudadanía crítica y democrática (PID2020-113453RB-I00), funded by the Agencia Estatal de Investigación (AEI/10.13039/501100011033). This paper has been thanks to the projects: PDC2022-133041-I00, funded by MCIN/AEI/10.13039/501100011033 and NextGeneration EU./PRTR, the project PID2020-113453RB-I00 funded by MCIN/AEI/10.13039/501100011033, the project 2020-1-ES01-KA226-HE-095430, funded by SEPIE in the call ERASMUS+ KA226, and "Ayuda de Recualificación del Profesorado Universitario", funded by Ministry of Universities and EU-NextGenerationEU.

## Conflict of interest

The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

## Publisher's note

All claims expressed in this article are solely those of the authors and do not necessarily represent those of their affiliated organizations, or those of the publisher, the editors and the reviewers. Any product that may be evaluated in this article, or claim that may be made by its manufacturer, is not guaranteed or endorsed by the publisher.

## Supplementary material

The Supplementary Material for this article can be found online at: https://www.frontiersin.org/articles/10.3389/feduc.2023.1213358/full#supplementary-material

## References

Ayre, C., and Scally, A. J. (2014). Critical values for Lawshe's content validity ratio: revisiting the original methods of calculation. *Meas. Eval. Counsel. Dev.* 47, 79–86. doi: 10.1177/0748175613513808

Bangdiwala, S. (1987). Using SAS software graphical procedures for the observer agreement chart. *Proc. SAS Users Group Int. Conf.* 12, 1083–1088.

Carretero, M., and López, C. (2009). Estudios cognitivos sobre el conocimiento histórico. Aportaciones para la enseñanza y alfabetización histórica. *Enseñanza Ciencias Soc.* 8, 75–89.

Carretero, M., and Van Alphen, F. (2014). Do master narratives change among high school students? A characterization of how national history is represented. *Cogn. Instruct.* 32, 290–312. doi: 10.1080/07370008.2014.919298

Castejón, J. L., Gilar, R., and Sánchez, B. (2009). "Modelos de enseñanza-aprendizaje," in *Aprendizaje, desarrollo y disfunciones: implicaciones para la enseñanza en la Educación Secundaria*, eds J. L. Castejón, and L. Navas (Alicante: Editorial Club Universitario), 7–48.

Chapman, A. (2011). Taking the perspective of the other seriously? Understanding historical argument. *Educ. Rev.* 42, 95–106. doi: 10.1590/S0104-40602011000500007

Cicchetti, D. V., and Feinstein, A. R. (1990). High agreement but low Kappa: II. Resolving the paradoxes. *J. Clin. Epidemiol.* 43, 551–558. doi: 10.1016/0895-4356(90)90159-M

Colomer, J. C., Sáiz, J., and Valls, R. (2018). Competencias históricas y actividades con recursos tecnológicos en libros de texto de Historia: nuevos materiales y viejas rutinas. Ensayos. *Rev. Fac. Alb.* 33, 1.

Danielson, C. (1996). *Enhancing professional practice: a framework for teaching.* Alexandria, VA: Association for Supervision and Curriculum Development.

De Graaff, R., Koopman, G., Anikina, Y., and Westhoff, G. J. (2007). An observation tool for effective L2 pedagogy in Content and Language Integrated Learning (CLIL). *Int. J. Biling Educ. Biling* 10, 603–624. doi: 10.2167/beb462.0

Desimone, L. M. (2009). Improving impact studies of teachers' professional development: toward better conceptualizations and measures. *Educ. Res.* 38, 181–199. doi: 10.3102/0013189X08331140

Domínguez Castillo, J. (2015). *Pensamiento histórico y evaluación de competencias.* Barcelona: Graó.

Feinstien, A. R., and Cicchetti, D. V. (1990). High agreement but low Kappa: I. The Problems of two paradoxes. *J. Clin. Epidemiol.* 43, 543–549. doi: 10.1016/0895-4356(90)90158-L

Fleiss, J. L., and Cohen, J. (1973). The equivalence of weighted kappa and the intraclass correlation coefficient as measures of reliability. *Educ. Psychol. Meas.* 33, 613–619. doi: 10.1177/001316447303300309

Friendly, M., and Meyer, D. (2016). *Discrete Data Analysis With R: Visualization and Modeling Techniques for Categorical and Count Data*. Chapman & Hall. Available online at: https://www.routledge.com/Discrete-Data-Analysis-with-R-Visualization-and-Modeling-Techniques-for/Friendly-Meyer/p/book/9781498725835

Galvis, R. (2007). De un perfil docente tradicional a un perfil docente basado en competencias. *Acc. Pedag.* 16, 48–57.

Gertsen, R., Baker, S. K., Haager, D., and Graves, A. W. (2005). Exploring the role of teacher quality in predicting the reading out comes for first-grade English learners: an observational study. *Remed. Spec. Educ.* 26, 197–206. doi: 10.1177/07419325050260040201

Gestsdóttir, S. M., van Boxtel, C., and van Drie, J. (2018). Teaching historical thinking and reasoning: construction of an observation instrument. *Br. Educ. Res. J.* 44, 960–981. doi: 10.1002/berj.3471

Gómez Carrasco, C. J., Chaparro Sáinz, A., Rodríguez-Medina, J., and Alonso-García, S. (2022). Recursos digitales y enfoques de enseñanza en la formación del profesorado de historia. *Edudación XXI* 25, 143–170. doi: 10.5944/educxx1.30483

Gómez Carrasco, C. J., Ortuño Molina, J., and Molina Puche, S. (2014). Aprender a pensar históricamente. Retos para la historia en el siglo XXI. *Temp. Arg.* 6, 5–27. doi: 10.5965/2175180306112014005

Gómez Carrasco, C. J., Rodríguez Medina, J., Miralles Martínez, P., and Arias González, V. B. (2020). Effects of a teacher training program on the motivation and satisfaction of History secondary students. *Rev. Psicod.* 26, 45–52. doi: 10.1016/j.psicoe.2020.08.001

Gómez Carrasco, C. J., Rodríguez Pérez, R. A., and Monteagudo Fernández, J. (2017). "Las competencias históricas en los procesos de evaluación: libros de texto y exámenes", in *Enseñanza de la historia y competencias educativas*, eds R. López, P. Miralles, J. Prats, and C. J. Gómez (Barcelona: Graó), 141–165.

Gómez Carrasco, C. J., Rodríguez-Medina, J., Miralles-Martínez, P., and López-Facal, R. (2021). Motivation and perceived learning of secondary education history students. Analysis of a programme on initial teacher training. *Front. Psychol.* 12, 661780. doi: 10.3389/fpsyg.2021.661780

Gómez, C. J., Monteagudo Fernández, J., and Miralles Martínez, P. (2018). Conocimiento histórico y evaluación de competencias en los exámenes de Educación Secundaria. Un análisis comparativo España-Inglaterra. *Educatio* 36, 85. doi: 10.6018/j/324181

Grossman, P., Loeb, S., Cohen, J., Hammerness, K., Wyckoff, J., Boyd, D., et al. (2010). *Measure for Measure: The Relationship Between Measures of Instructional Practice in Middle School English Language Arts and Teachers' Value-Added Scores (No. w16015).* Cambridge, MA: National Bureau of Economic Research.

Guerrero-Romera, C., Sánchez-Ibáñez, R., and Miralles-Martínez, P. (2022). Approaches to history teaching according to a structural equation model. *Front. Educ.* 7, 842977. doi: 10.3389/feduc.2022.842977

Gwet, K. L. (2008). Computing inter-rater reliability and its variance in the presence of high agreement. *Br. J. Math. Stat. Psychol.* 61, 29–48. doi: 10.1348/000711006X126600

Gwet, K. L. (2021). *Handbook of Inter-rater Reliability*. AgreeStat Analytics. Available online at: https://www.researchgate.net/publication/267922774_Handbook_of_inter-rater_reliability_The_definitive_guide_to_measuring_the_extent_of_agreement_among_raters

Henze, I., and van Driel, J. H. (2011). "Toward a more comprehensive way to capture PCK in its complexity," in *Re-Examining Pedagogical Content Knowledge*, eds A. Berry, P. Friedrichsen, and J. Loughran (New York, NY: Routlegde), 120–134.

Hernández, F., Maquilón, J. J., and Monroy, F. (2012). Estudio de los enfoques de enseñanza en profesorado de educación primaria. *Rev. Curric. Form. Prof.* 16, 61–77.

Hill, H. C., Charalambous, C. Y., and Kraft, M. (2012). When rater reliability is not enough: observational systems and a case for the G-study. *Educ. Res.* 41, 56–64. doi: 10.3102/0013189X12437203

Huijgen, T., Holthuis, P., Van Boxtel, C., and Van de Grift, W. (2019). Promoting historical contextualisation in classrooms: an observational study. *Educ. Stud.* 45, 456–479. doi: 10.1080/03055698.2018.1509771

Huijgen, T., Van de Grift, W., Van Boxtel, C., and Holthius, P. (2017). Teaching historical contextualization: the construction of a reliable observation instrument. *Eur. J. Psychol. Educ.* 32, 159–181. doi: 10.1007/s10212-016-0295-8

Kember, D., and Kwan, K. P. (2000). Lecturers' approaches to teaching and their relationship to conceptions of good teaching. *Instruct. Sci.* 28, 469–490. doi: 10.1023/A:1026569608656

Lawshe, C. H. (1975). A quantitative approach to content validity. *Pers. Psychol.* 28, 563–575. doi: 10.1111/j.1744-6570.1975.tb01393.x

López, R., and Valls, R. (2012). "La necesidad cívica de saber historia y geografía," in *Educar para la participación ciudadana en la enseñanza de las Ciencias Sociales*, eds N. de Alba, F. F. García, and A. Santisteban (Sevilla: Díada Editora), 185–192.

Matsumura, L., Garnier, H., Slater, S., and Boston, M. (2008). Toward measuring instructional interactions atscale. *Educ. Assess.* 13, 267–300. doi: 10.1080/10627190802602541

Mercer, N. (2010). The analysis of classroom talk: methods and methodologies. *Br. J. Educ. Psychol.* 80, 1–14. doi: 10.1348/000709909X479853

Monte-Sano, C., and Budano, C. (2013). Developing and enacting pedagogical content knowledge for teaching history: an exploration of two novice teachers' growth over three years. *J. Learn. Sci.* 22, 171–211. doi: 10.1080/10508406.2012.742016

Muñoz, S. R., and Bangdiwala, S. (1997). Interpretation of kappa and B statistics measures of agreement. *J. Appl. Stat.* 24, 105–112. doi: 10.1080/02664769723918

Nokes, J. D. (2010). Observing literacy practices in history classrooms. *Theory Res. Soc. Educ.* 38, 515–544. doi: 10.1080/00933104.2010.10473438

Oattes, H., Wilschut, A., Oostdam, R., Fukkink, R., and de Graaff, R. (2022). Practical solution or missed opportunity? The impact of language of instruction on Dutch history teachers' application of pedagogical content knowledge (PCK). *Teach. Teach. Educ.* 115, 103721. doi: 10.1016/j.tate.2022.103721

Olmos, R. (2017). Kahoot: ¡Un, dos, tres! Análisis de una aplicación de cuestionarios. *Íber Didác. Cien. Soci. Geo. His.* 86, 51–56.

Peck, C., and Seixas, P. (2008). Benchmarks of historical thinking. first steps. *Can. J. Educ.* 31, 1015–1038.

Pianta, R., Hamre, B., and Mintz, S. (2011). *Classroom Assessment Scoring System: Secondary Manual.* Charlottesville, VA: Teachstone,.

Pianta, R., La Paro, K., and Hamre, B. (2008). *Classroom assessment scoring system (CLASS).* Baltimore, MD: Paul H. Brookes.

R Core Team (2021). *R: A Language and Environment for Statistical Computing.* Vienna: R Foundation for Statistical Computing. Available online at: https://www.R-project.org/ (accessed April 15, 2023).

Sáiz Serrano, J., and Gómez Carrasco, C. J. (2016). Investigar pensamiento histórico y narrativo en la formación del profesorado: fundamentos teóricos y metodológicos. *Rev. Electrón. Int. Form. Prof.* 19, 175–190.

Sánchez, R., Campillo, J. M., and Guerrero, C. (2020). Percepciones del profesorado de primaria y secundaria sobre la enseñanza de la historia. *Rev. Int. Form. Del Prof.* 34, 57–76. doi: 10.47553/rifop.v34i3.83247

Santiesteban Fernández, A. (2010). La formación de competencias de pensamiento histórico. *Clío Asoc.* 14, 34–56. doi: 10.14409/cya.v1i14.1674

Sawada, D., Piburn, M., Judson, E., Turley, J., Falconer, K., Benford, R., et al. (2002). Measuring reform practices in science and mathematics classrooms: the reformed teaching observation protocol. *School Sci. Math.* 102, 245–253. doi: 10.1111/j.1949-8594.2002.tb17883.x

Schoenfeld, A. (2013). Classroom observations in theory and practice. *ZDM* 45, 607–621. doi: 10.1007/s11858-012-0483-1

Seixas, P., and Morton, T. (2013). *The Big Six Historical Thinking Concepts*. Toronto, ON: Nelson College Indigenous.

Sobejano, M. J., and Torres, P. A. (2009). *Enseñanza de la historia en Secundaria: historia para el presente y la educación ciudadana*. Madrid: Tecnos.

Stallings, J., and Kaskowitz, D. (1974). *Follow Through Classroom Observation Evaluation, 1972–1973.* Menlo Park, CA: Stanford Research Institute.

Stearns, L., Morgan, J., Capraro, M., and Capraro, R. (2012). A teacher observation instrument for PBL classroom instruction. *J. STEM Educ. Innovat. Res.* 13, 7–16.

Teddlie, C., Creemers, B., Kyriakides, L., Muijs, D., and Yu, F. (2006). The international system for teacher observation and feedback: an evolution of an international study of teacher effectiveness constructs. *Educ. Res. Eval.* 12, 561–582. doi: 10.1080/13803610600874067

Trigwell, K., and Prosser, M. (2004). Development and use of the approaches to teaching inventory. *Educ. Psychol. Rev.* 16, 409–424. doi: 10.1007/s10648-004-0007-9

Trigwell, K., Prosser, M., and Ginns, P. (2005). Phenomenographic pedagogy and a revised approaches to teaching inventory. *Higher Educ. Res. Dev.* 24, 349–360. doi: 10.1080/07294360500284730

Valls, R., and López, R. (2011). "Un nuevo paradigma para la enseñanza de la historia? Los problemas reales y las polémicas interesadas al respecto en España y en el contexto del mundo Occidental. *Ens. Las Cien. Soc.* 10, 75–85.

Van Boxtel, C., van Drie, J., and Stoel, G. (2020). "Improving teachers' proficiency in teaching historical thinking," in *The Palgrave Handbook of History and Social Studies Education*, eds C. W. Berg, and T. M. Christou (Springer International Publishing), 97–117. Available online at: https://link.springer.com/chapter/10.1007/978-3-030-37210-1_5

Van de Grift, W. (2007). Quality of teaching in four European countries: a review of the literature and an application of an assessment instrument. *Educ. Res.* 49, 127–152. doi: 10.1080/00131880701369651

Van de Grift, W. (2009). Reliability and validity in measuring the value added of schools. *School Eff. Sch. Improv.* 20, 269–285. doi: 10.1080/09243450902883946

Van Drie, J., and van Boxtel, C. (2008). Historical reasoning: towards a framework for analyzing students' reasoning about the past. *Educ. Psychol. Rev.* 20, 87–110. doi: 10.1007/s10648-007-9056-1

Van Hover, S., Hicks, D., and Cotton, S. (2012). Can you make historiography sound more friendly? Teaching observation instrument. *Hist. Teach.* 45, 603–612.

Vermunt, J. D., and Verloop, N. (1999). Congruence and friction between learning and teaching. *Learn. Instruct.* 9, 257–280. doi: 10.1016/S0959-4752(98)00028-0

Wineburg, S. (2001). *Historical Thinking and Other Unnatural Acts. Charting the Future of Teaching the Past.* Philadephia, PA: Temple University Press.