HISTORICAL ENCOUNTERS
A journal of historical consciousness, historical cultures, and history education

ISSN 2203-7543 | Double Blind Peer-Reviewed | Open Access | www.hej-hermes.net

A competence-based test to assess historical thinking in secondary education: Design, application, and validation

Jesús Domínguez-Castillo
Universidad Antonio de Nebrija, Spain

Laura Arias-Ferrer
Universidad de Murcia, Spain

Raquel Sánchez-Ibáñez
Universidad de Murcia, Spain

Alejandro Egea-Vivancos
Universidad de Murcia, Spain

Francisco Javier García-Crespo
Universidad Complutense de Madrid, Spain

Pedro Miralles-Martínez
Universidad de Murcia, Spain

ABSTRACT

This paper presents the theoretical framework, application and final outcomes of a pilot test designed as a possible model for assessing students' historical thinking in Secondary Education. It is based both on widely accepted historical thinking concepts and on the assessment framework developed by PISA. The test tries to assess what could be named as the three major competences in history: "explain historically", "use of sources as historical evidence" and "understanding the features of historical knowledge". It includes several stimuli (texts, images…) and a total of 39 items. The field trial of the test was applied to a convenience sample of 893 10th and 11th grade students, aged 16 to 18 years. Their answers were analysed statistically according to the Item Response Theory (IRT), and the results uphold the validity and reliability of the test instrument. The IRT analysis also enables us to take a first step towards defining levels of achievement and progress for the learning and acquisition of those competences. One implication of note of this research is the possible adoption of this model for assessing history, based both on applied content knowledge and historical thinking concepts and skills. Such a model of assessment would also stimulate more active, problem-based and motivating teaching approaches.

KEYWORDS

Competencies, Historical thinking, Assessment, Pilot study, Item Response Theory (IRT), Educational research

CITATION

Domínguez-Castillo, J., Arias-Ferrer, L., Sánchez-Ibáñez, R., Egea-Vivancos, A., García-Crespo, F. J., & Miralles-Martínez, P. (2021). A competence-based test to assess historical thinking in secondary education: Design, application, and validation. *Historical Encounters, 8*(1), 30-45. https://doi.org/10.52289/hej8.103

Introduction

The incorporation of key competences into education curricula as a recommendation of the European Union since 2006 [1] has led to reflection and debate about the meaning of these competences and their transfer to the classroom. The task is not proving easy in Spanish Secondary Education, where subjects remain the backbone of syllabuses and no detailed indications have been provided regarding the effects that the eight European key competences should have on the curricula. Fortunately, international studies of recognized prestige, like the OECD's PISA and, to a lesser extent, PIRLS and TIMSS of the *International Association for the Evaluation of Educational Achievement* (IEA), have explained very clearly the meaning that competences in education have for them. PISA, in particular, is very explicit in this respect and the opening lines of its theoretical framework of 2015 read (OECD, 2017):

> PISA assesses the extent to which 15-year-old students, near the end of their compulsory education, have acquired key knowledge and skills that are essential for full participation in modern societies. The triennial assessment focuses on the core school subjects of Science, Reading and Mathematics. […] The assessment does not just ascertain whether students can reproduce knowledge; it also examines how well students can extrapolate from what they have learned and can apply that knowledge in unfamiliar contexts, both in and outside school. (p. 12)

Two ideas from the above are worth highlighting. The first, that competence is having knowledge and being able to apply it to new situations both inside and outside the school; and the second, that PISA does not assess new knowledge or skills proper to one or another European key competences, such as "maths and science competence", "digital competence", etc., but those of the traditional subjects (Maths, History, etc.) although with a functional or applied focus on what is learnt. This is why we prefer to speak here about *competence-based assessment of history* instead of the *assessment of key competences in history*.

It is clear that international assessment studies have oriented Education administrators and teachers in terms of what competences can bring to Maths and Science education. But this is not the case for History, with no model of competence-based assessment. In order to cover this gap, we have tried to build up a tentative proposal for this model of assessment in History, implemented and tested in Spanish education contexts. Our starting point was the theoretical framework used by the OECD in the assessment of Sciences which, as with other subjects, always comprises three elements: "situations or contexts", i.e., facts or cases in which to apply knowledge; "content knowledge" or contents of the subject; and "competences or processes". The latter constitute the essential cognitive strategies, specific to each subject discipline, that students have to use to address and answer the questions and problems that often arise in daily life. Hence, these competences are the key element in the PISA test and the driving force behind the choice and design of items (OECD, 2017, p. 42).

In the case of History, and using the elements defined in PISA framework, "situations or contexts" can mean unfamiliar past or present events or problems, for which historical knowledge

(and skills) is worth extrapolating to understand and address them. As for the "content knowledge" of History, it usually includes facts, people and processes in human past, that school curricula generally present under "substantive concepts" (i.e. Reformation, Enlightenment, Industrial Revolution, etc.). But, which are the "competences" or "processes" specific to the discipline of history? To address this question, we draw on research into History education and particularly, historical thinking. Authors insist on the idea that to think historically is not only to know what happened in the past but also "to understand how knowledge has been constructed and what it means" (Lévesque, 2008, p. 27). At the core of historical thinking, authors put forward a series of concepts and skills of historical methodology (evidence, causes, change and continuity, accounts, etc.) as being essential for understanding what scientific knowledge of our past is (Lee, 1983 and 2005; Lee & Shemilt, 2003; Lévesque, 2008; Seixas, 1996; Seixas & Peck, 2004; Seixas & Morton, 2013). Such meta-historical or methodological contents have been named "second order" concepts to distinguish them from the previously mentioned substantive concepts (or the so-called "first order" concepts). They represent the core features of history as a particular discipline of scientific knowledge: For instance, historians relay on critical analysis of evidence (sources and traces) to know about human past, while natural scientists use experimental tests to learn about the natural world. These type of concepts need to be taught mainly through skills practice, complemented by some reflection on the historical way of knowledge.

Recent years have also seen publications which adopt various angles in their search for how to assess this historical thinking (González, Pagès & Santisteban, 2011; Eliasson, Alvén, Yngvéus & Rosenhud, 2015; Körber & Meyer-Hamme, 2015; Seixas, Gibson & Ercikan, 2015; VanSledright, 2014; Wilschut, 2015; Wineburg & Steinberg, 2007). Several features can be highlighted from their proposals. First of all, the central role that historical thinking concepts play in their assessment framework: Items trying to assess evidence, causation, historical perspective or empathy, and change and continuity, are commonly included among others like consciousness or ethical dimension. Many of these tests are also based on primary sources as stimuli, presenting some "unknown" events to be analyzed and 'solved' by students (Duquette, 2015; Eliasson et al., 2015; Körber & Meyer-Hamme, 2015; Seixas, Gibson, & Ercikan, 2015). In this line, our test was mainly focused on two particular topics referring to historical periods recently studied by students: Child labour during the Industrial Revolution and Spanish migration during Franco's Dictatorship (Domínguez, 2015; Domínguez, Arias, Sánchez, Egea, & García, 2016; Sáiz & Fuster, 2014). This allowed us to assess skills of historical competences but also the ability to apply contextual knowledge of the period (Arias et al., 2019). As to the type of assessment items, we choose to include multiple-choice and short answer questions instead of long historical accounts (frequent in the previous proposals).

Based on the above literature we have tried to adapt the PISA model to propose a theoretical framework and an assessment for History based on historical thinking concepts. It is here wherein lies the originality of this research, the aim of which is reflected in the two research questions that make up this paper: Is it possible to design a valid and reliable test to assess students' historical thinking when they finish compulsory education? Do the results of our research afford a basis on which to build a scale of the progress in the acquisition of competences and concepts of historical thinking?

Method

In accordance with the two guiding aims of our research, this pilot study seeks to assess the suitability of the test and not the historical knowledge and skills of the participating students.

There are three clearly defined stages within our research: Assessment framework and design of the test, trialling the test through fieldwork, and statistical analysis of the responses.

Assessment framework and design of the test[2]

Table 1 summarizes the assessment framework of our test, which has been explained in greater detail in earlier publications (Domínguez, 2015; Domínguez et al., 2016). Following the PISA model for sciences (OECD, 2017 and previous editions), three large "historical competences" are defined which form the basis of the test. To these are fitted the items or questions that assess the substantive knowledge and the skills in which the "second order" historical concepts are embodied and which are distributed in similar proportions to those used in the sciences in PISA.

Explain (past and present) facts historically is the equivalent competence in Sciences as "Explaining phenomena scientifically" (OECD, 2017, p. 21). It is directly focused towards assessing the capacity to apply historical substantive knowledge to the facts or problems posed in the questions. For example, to answer questions about emigration in Spain during the 1950s and 1960s it is necessary to have some minimum contextual knowledge of the Spanish Civil War and the Franco Regime. This competence includes placing the events studied within their time framework and context, indicating background events, and even identifying notable historical events.

Competences or cognitive processes of history	**Type of knowledge**	**"Second order" concepts**	**Number and percentage of items**
Explain historically	Substantive knowledge	--	14 (36%)
Use historical evidence	Methodological knowledge	"Sources" and "evidence"	13 (33.3%)
Understand the features of historical knowledge	Methodological knowledge	"Causation", "Empathy, "Change and continuity"…	12 (30.7%)

Table 1. Assessment framework and distribution of items according to competences, type of knowledge and second order concepts of history

The two other competences -*Use historical evidence* and *Understand the features of historical knowledge*– have their equivalents in PISA 2015 Science framework– "Interpreting data and evidence scientifically" and "Evaluating and designing scientific enquiry" (OECD, 2017, p. 22). In both subjects, the two competences are directly related to methodological knowledge, that is, to the characteristic methods and skills of both disciplines. In History, these embrace the "second order" concepts of historical thinking, from which we have selected the four most widely accepted concepts to date by experts in History education (see Table 1). *Use historical evidence* is the concept of most weight in the test, as befits its importance in historiographic methodology. Besides, the competence *Understand the features of historical knowledge* includes other concepts that would form part of what some specialists have called "historical perspective" (Duquette, 2015, p. 52).

When assessing these concepts, it has been necessary to distinguish specific cognitive skills of each concept, as is shown in Table 2 for *Use of historical evidence.* This list of skills, which was drawn up from a review of the research and innovation in the teaching and learning of each concept (Domínguez, 2015), serves as a bridge between historical "second order" concepts and the items, so focusing the design of the tasks and the questions to evaluate each skill. Thus, in the case of the competence "Use of historical evidence", skills contemplate two major groups of skills: on the one hand, literacy skills, in order to get information from, and understand different source material and, on the other, specific skills related to the historians´ critical use of sources as evidence from the past. Both groups of skills are interconnected and need to be assessed by complementary test items.

The design phase took place in 2014 to 2015 and included qualitative validation by a diversified group of experts and a pre-pilot run with 49 students. The test was designed for

students in their final year of Compulsory Secondary Education (15-16 years old). We took into consideration the content-knowledge of history studied that year (World and Spanish history during the nineteenth and twentieth centuries), as well as some historical thinking skills included in the curriculum, unfortunately, without any rationale or justification. Having said that, the test was not designed to evaluate the outcomes of the Spanish curriculum in History. Certainly, it was assumed as the basis of students' knowledge, but our aim was to pilot a type of test able to assess students' historical thinking.

The test comprises two different assessment units: "Spanish emigrants yesterday and today (1950-2014)" (coded as M1), and "Child labor during the Industrial Revolution and today" (coded as M2). Both topics were chosen because they refer to historical periods included in the current curriculum and are particularly motivating and useful to understand the relation between past and present. Each unit included two booklets, one with the stimulus documents and another with the questions. The stimuli are in the main primary sources (texts, images, graphics, and maps) that the students should use, along with their own knowledge, to answer the questions. There were 39 questions in all: 18 were multiple choice, 13 were short-answer questions and 8 were semi-open. The questions were distributed according to the three competences shown in Table 1, and in approximately equal proportion.

Two groups of skills	**Detailed skills of "using evidence"**
Reading and communication skills	Obtain explicit information Infer information Integrate and interpret Assess information Synthesize and communicate
Historiographic skills	Contextualize sources Analyze them critically Collate and evaluate sources

Table 2. Skills for assessing the competence "Use historical evidence" (Domínguez, 2015, p. 90)

Trialling the test through fieldwork

The pre-pilot test led to some questions being reworded or even replaced, to the refinement of the estimated times to complete the test and to fine-tuning the application protocol and the coding rubrics.

The definitive field study was carried out in April and May in 2015 and 2016 using a convenience sample with an acceptable male-female split, venue, and socio-cultural level. To have a sufficiently large data base and so meet the IRT requirements, 893 students, aged 15 to 18 years, took part. Of these, 789 were in the fourth year of compulsory secondary education at ten state and private centres in Región de Murcia (Spain) and 109 in the first year of *Baccalaureate* in four state centres in Galicia (Spain). Each assessment unit was carried out in a class period of 45 minutes. Four people administered the tests, and, along with a fifth person, they scored and coded the booklets following the rubrics previously revised and agreed upon. Many tests were double-scored to ensure that any doubts were discussed and resolved in an agreed upon manner. As the statistical analysis would show, there was a good homogeneity and coherence in the administration, scoring and coding of the test.

Statistical analysis of the responses

After coding and recording the responses, the statistical analysis was performed based on the Item Response Theory (IRT) and using ConQuest©.

The IRT analysis simultaneously estimates the degree of difficulty and the students' skill. The level of difficulty of each question is determined by the probability of a correct response given each individual's level of skill. Likewise, the skill of each student is measured by taking into account the probability of a correct answer given the difficulty of the question. Therefore, the answer to one item depends on the interaction of the student's "skill" and its difficulty. This is why this analysis cannot show scores that are linked to a particular population or to a group of specific individuals, but scores that are based on the above relation (Embretson & Reise, 2000).

Results and discussion

Of the large number of data obtained with the ITR analysis, three points fundamental to our research are worth highlighting: the degree of discrimination of each item, the overall reliability coefficient of the test, and the distribution of items and students according to levels of difficulty and scores. The first two of these respond to the first aim of our research and the third to the second aim.[3]

Discrimination of each item

It discerns whether an item coherently distinguishes between students according of their skill in the test as a whole. Hence, a very simple question should be correctly answered by a large majority of the students, while a difficult one will be correctly answered by those showing a certain degree of skill. Three items (M1_8.1, M2_3 and M2_13) were removed due to bad discrimination, bad fit and bad behaviour of the distractor, which meant that the final test had 36 items. Let us look at M1_8.1 as an example of a rejected item (see Table 3).

Item M1_8.1			
Cases for this item	604	Discrimination .12	
Item Threshold(s):	.03	Weighted MNSQ	1.12
Item Delta(s):	.03		
Label	**Score**	**Count**	**% of total**
9	.00	26	4.30
A	.00	203	33.61
B	.00	89	14.74
C	1.00	257	42.55
D	.00	29	4.80

Table 3. Example of a rejected item: Item M1_8.1 (Emigrants unit, question 8.1)

The table shows a low discrimination .12 for an item of medium difficulty (.03) whose weighted MNSQ is 1.12, i.e., far from 1. All this bad behaviour is fundamentally because option "A" (distractor) was chosen almost as often as the correct option, "C". This in itself would not be a problem were it not for its having been chosen by a large number of students of all levels of skill.

The reliability coefficient

This is used to ascertain the stability of the test when obtaining results. It is calculated using the classic concept of dividing the real variance by the observed variance. In this case the real variance is an estimation of the estimation a posteriori (EAP) of the distribution, and the observed variance is an estimation obtained from the variance explained by the plausible values (PV). Once the 36 items have been definitively configured, the ensuing test shows a high coefficient, EAP/PV= .759.

To illustrate the calibration proper to each item, Table 4 shows the following parameters for the first eight items: estimation of difficulty (where 0 represents an average difficulty, a negative value means low difficulty and a positive value refers to high difficulty); the estimation error; the weighted MNSQ correction (which should be as close as possible to 1) and the confidence interval.

Item	Difficulty	Estimation error	Weighted MNSQ	Confidence Interval
M1_1	.306	.063	1.04	(.94- 1.06)
M1_2	.607	.064	1.01	(.93-1.07)
M1_3.1	-2.326	.072	.98	(.84-1.16)
M1_3.2	-.386	.053	1.08	(.90-1.10)
M1_4	.893	.065	.98	(.91-1.09)
M1_5.1	-.993	.063	.97	C
M1_5.2	.921	.066	1	(.91-1.09)
M1_6.1	-1.796	.068	.98	(.88-1.12)

Table 4. Example of the calibration of some items

Next, we give two examples of analyses of items showing good behaviour: The first (Item 28, M2_5) is multiple choice and is coded 0-1, and the second (Item 31, M2_7.2) is an open question and is coded 0-1-2.

Item 28 (M2_5):

*Choose among the following options the **only** one that points out the most important difference between cottage industry and mechanized industry:*

a) Industrial machines made it more difficult the workers' tasks

b) Mechanical industry has iron machines, while cottage industry has them made of wood.

c) Industrial machines multiplied artisans' production (correct answer)

d) In cottage industry each machine needed a person while in mechanical industry it needed several persons

Item 31 (M2_7.2):

Was the scavenger' work dangerous for kids? Docs. 5 and 6 do not agree on that.

a) What do they assert each of them?

b) Which assertion is best supported by other documents?

The basic parameters are given in Tables 5 and 6, and the characteristic curves are shown in Figures 1, 2 and 3). Item 28 was answered by a sufficiently large number of students (582), of which 398 (63.38%) answered correctly. IRT analysis shows that it is of low difficulty (-1.18, with an estimation error of .064), which means that a student with skill -1.18 has a .5 probability of

answering the question correctly. From Figure 1, which relates the student's skill and the probability of a correct answer, it is also inferred that a student of skill 1 has a .9 probability. The discrimination value of this item is .37, which is reasonable. Finally, the "weighted MNSQ", which indicates the relation between the item and the overall test, is exactly 1. All these factors allow us to conclude that the item behaves well.

In order to complete the analysis of item 28 (M2_5) we point to Figure 2, which shows the characteristic curve of the item in relation to the five possible response categories (9, A, B, C and D). The correct answer is "C" (the continuous line with crosses) and is the option that is chosen more frequently as the skill of the student increases. Option "D" (the broken line with circles) is the second most answered option, and it decreases with the students' skill (the greater the skill the lower the probability of its being the answer given). The other options remain constant, with values very close to zero.

Item 28 (M2_5)
Cases for this item 582 Discrimination .37
Item Threshold(s): -1.18 Weighted MNSQ 1.00
Item Delta(s): -1.18

Label	**Score**	**Count**	**% of total**
9	.00	14	2.41
A	.00	31	5.33
B	.00	24	4.12
C	1.00	398	68.38
D	.00	115	19.76

Table 5. Example 1 of the analysis of an item that behaves well: 28 (M2_5)

Item 31 was also answered by a sufficient number of students (582), of which 190 (32.65%) answered it partially correctly (code 1), 76 (13.06%) totally correctly (code 2), 259 (44.50%) badly and 57 (9.79%) left it blank (Table 6). The item is of medium-high difficulty since, its thresholds are -.13 and 1.25 for the answers scoring 0 and 2, respectively. Thus, it means that a student with a skill -.13 has a .5 probability of answering the item incorrectly and a student with a skill 1.25 has a .5 probability of scoring the maximum. The discriminatory value of this item is .47, which is within reasonable parameters. The weighted MNSQ is .99, practically 1. All of this points to its being an item that *behaves well* (Figure 3).

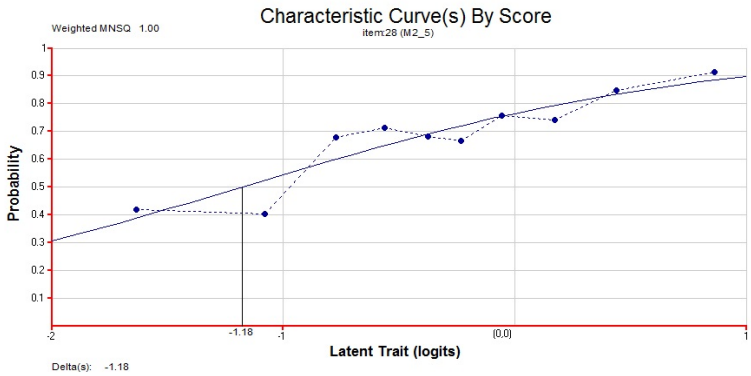

Figure 1. Characteristic curve of the item 28 (M2_5)

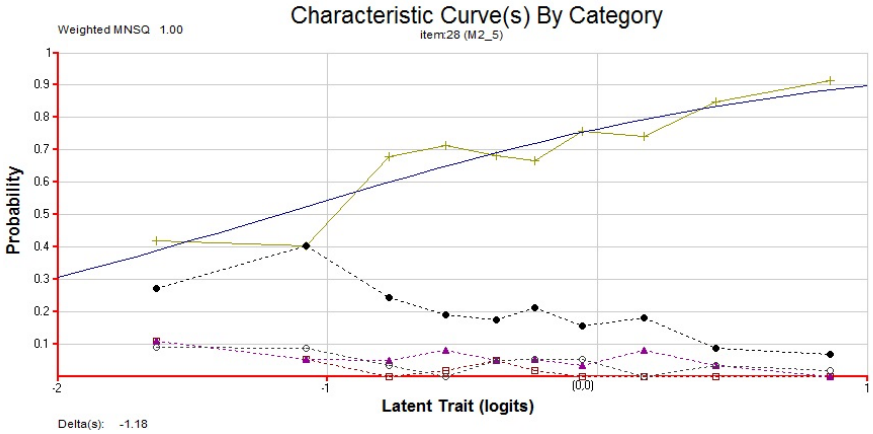

Figure 2. Characteristic curve of item 28 (M2_5) by category

Item 31 (M2_7.2) Cases for this item 582 Discrimination .47 Item Threshold(s): -.13 1.25 Weighted MNSQ .99 Item Delta(s): .16 .97			
Label	**Score**	**Count**	**% of total**
0	.00	259	44.50
1	1.00	190	32.65
2	2.00	76	13.06
9	.00	57	9.79

Table 6. Example 2 of the analysis of an item showing good behaviour: 31 (M2_7.2)

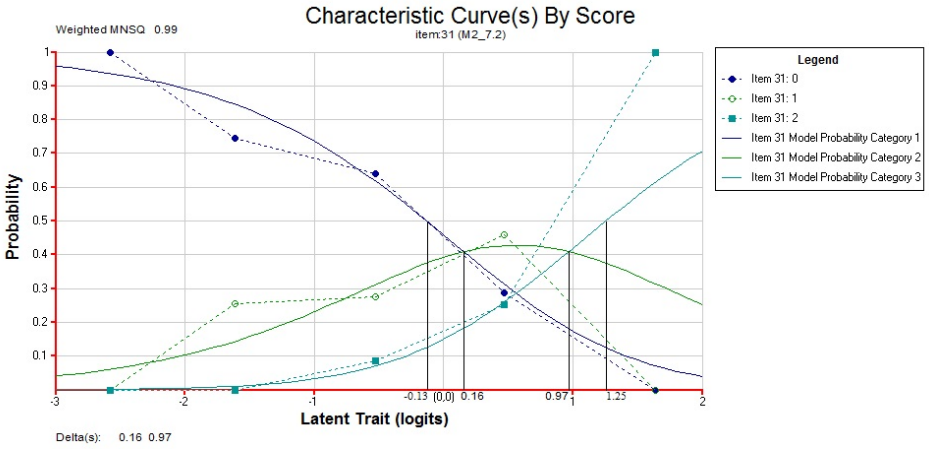

Figure 3. Characteristic curve of item 31 (M2_7.2)

Distribution of items and students by score

This is the third of the key data returned by the IRT. It uses a single scale to show the degree of difficulty of the items and the scores obtained by the students. Figure 4 shows the graph of the test: the vertical scale is a difficulty skill scale from -2 to 2, and the items are distributed on the right (indicated by their number, including the code, 1 or 2) and to the left of these, in the centre column, is the number of students' answers in each level of skill (each "x" represents 5.2 cases). The figure shows that the test has a very balanced distribution of items and answers, although the deduction is that, overall, the test was difficult as the students showed a negative mean score for skill, while, in contrast, the items are accumulated above 0 and are therefore positive, and harder.

As occurs in most large-scale assessment studies (PISA, TIMSS, PIRLS, etc.), the above scoring scale is usually modified to remove decimals and so facilitate its reading. This is done by shifting the middle point from 0 to 500, point 1 to 600, point 2 to 700, etc. This measure also has a standard deviation of 100, which means that approximately two-thirds of the students scored between 400 and 600 points, and 95% scored between 300 and 700 points. With this scale the test items are also distributed according to their degree of difficulty, from 273 points for the easiest item (item 3, M1_3.1) to 785.74 for the most difficult (item 23.2, M2_1.2-.2, i.e., coded 2).

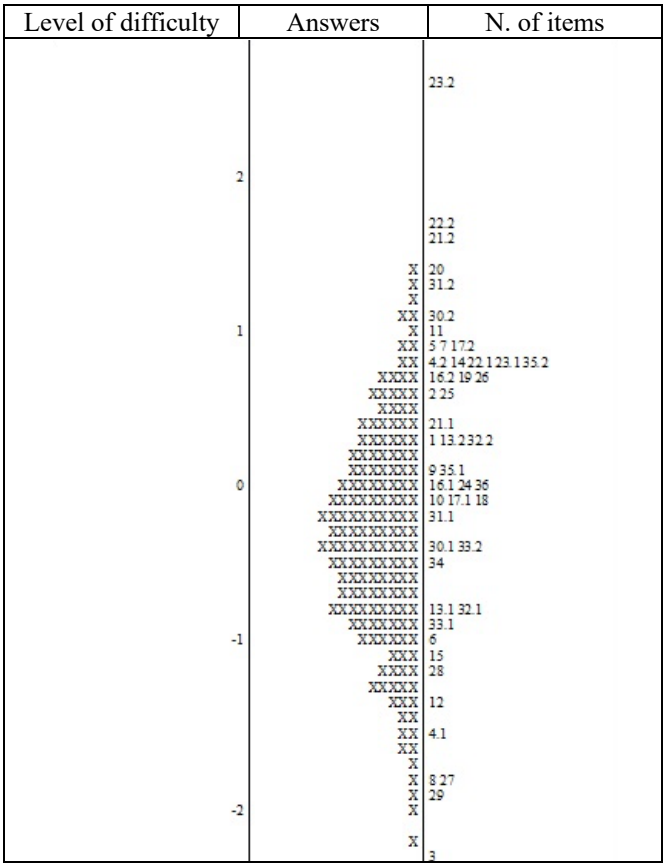

Figure 4. Distribution of items and responses according to the level of difficulty

Towards a tentative scale of progression in historical thinking

In line with the previously mentioned large scale assessment studies, within this range of scores we distinguish five levels of performance, according to the items students were able to answer correctly. The lowest level is level 1, which is characterized by the fact that it is not possible to know the skills of those who answered only these items correctly. Its upper limit was established, in our test, at 392 points. From there on each successive level is every 64 points, up to the fifth,

with the highest score of 785.74 points. A clarification is in order here: the test contains 12 questions with marking codes 0, 1 and 2, assigned according to the accuracy and quality of the answers. This enables us to fine tune the levels of the answers, albeit that IRT provides two scores for these items, one for answers coded 1 and another, higher one, for answers that are coded 2. Thus, as Table 7 shows, for the effects of the analysis and the description of the possible levels of responses, we have 48 item-responses (36 questions plus 12 responses in questions with code 2).

Number of items and item-responses per evaluation unit	**Use historical evidence**	**Causal explanation**	**Empathetic explanation**	**Time, change, continuity**	**Explain historically**
M1 unit: 21/26	6/7	--	3/3	4/7	8/9
M2 unit: 15/22	6/10	4/6	--	1/2	6/4
Total items: 36 Item-Responses: 48	Items: 12 Items-Res: 17	Items: 4 Items-Res: 6	Items: 3 Items-Res: 3	Items: 5 Items-Res: 9	Items: 14 Items-Res: 13

Table 7. Number of items and item-responses (Item-Res)
by competence and methodological concept of history

Level	**Value**	**Descriptors**
1	273-392	Four items appear at this level. It is worth stating that students partially analyze and interpret very basic textual or graphic information with few details; they make very simple inferences and interpretations that are partially contextualized and draw on their elementary historical knowledge (e.g., the use of the steam engine in the United Kingdom).
2	392-456	These items analyze basic information (text, illustrations and maps) with limited detail; they make simple contextualized inferences and interpretations drawing on limited historical knowledge (e.g. about the precariousness of agriculture in the post Spanish Civil War period) and geographical knowledge (they partially identify large countries on a blank political map); they recognize that sources can have differing values or uses; they collate, and partially identify, with mistakes, the information afforded by sources about the work of the "scavenger" girl.
3	456-520	They analyze and interpret correctly a bar chart about migratory figures in Spain (1960-2014), distinguishing stages and their overall significance. They collate and distinguish the information obtained from two sources about the work of the "scavenger" girl, but they are not able to evaluate appropriately which of the two offers the best founded version (Item 31.1/M2_ 7.2.1 score 1out of 2: 476,9 points)
4	520-584	They analyze various (2 to 4) documents in detail and obtain explicit information, they make precise inferences, drawing at times on contextualization ("the steam engine that probably moves the spinning machine"), collate, integrate and interpret information from several documents to detect errors or contradictions (date of the grandparents' wedding) or to obtain proofs on which to ground certain statements (that the great grandfather was a very small landowner), and they partially synthesize information obtained from various sources (transformations in Spain from 1960 to the present day).
5	>584	They analyze rigorously and collate in detail several sources to obtain explicit information about facts (the work of the scavenging girl); collate two sources and evaluate which offers the version best supported by the other documents (whether "scavenging" was dangerous for children's health) (Item 31/M2_7.2.2 score 2: 606,5 points). Finally, they synthesize and communicate with precision the information supplied by various sources about the changes to Spanish society.

Table 8. Levels of performance associated to the competence 'using sources as historical evidence'

From the scores for these 48 item-responses it is possible to describe and characterize the levels of success in historical competences and, where appropriate, those in the methodological or 'second order concepts' of the discipline. In two of the three competences evaluated – 'Explain events historically' and 'Use historical evidence,' it is possible to characterize the majority of the five levels since they have 13 and 17 item-responses, respectively. In the case of the third competence –'Understand the logic of historical knowledge'- it is not possible to distinguish the response levels with rigor. On the one hand, the items associated with this competence evaluate three different concepts in themselves (causation, empathy, and change and continuity), with just 6, 3 and 9 item-responses for each one. On the other hand, each concept requires a specific scale that is independent of those of the other concepts and that does not allow it to establish level equivalences between, for example, the students' achievement in causal explanation and in change and continuity.

Tables 8 and 9 present an initial classification and description of the levels of performance for the two competences mentioned, thus showing the possibilities that these types of tests afford research into the progression in our students' learning. Notwithstanding the test's limitations, this first scale in the progression of responses opens up a route for future research that may fine tune even further the different levels of response of our students in competences and historical methodology concepts. The ensuing results will serve to enrich the already important empirical base available (Lee & Shemilt, 2003; Seixas & Morton, 2013).

Level	**Value**	**Descriptors**
1	273-392	Not enough data to characterize this level. There are only two items in this level (28/M2_5 and 29/M2_6), and both are multiple choice (identify differences between crafts and industry).
2	392-456	Not enough data. There is only one item of this difficulty (34/M2_10), which is again multiple choice.
3	456-520	Scarce data. Most items are multiple choice, so identification is predominant. Some historical knowledge is noticeable and the students connect it to the context of the topics in the test: there is an acceptable knowledge of chronology in Spain from the Civil War to the present (1/M1_1), and of the significance of the rural exodus in Spain under the Franco regime (10/M1_7.1); in M2, students identify that colonization is partly responsible for delayed economic development in many countries (34/M2_10),
4	520-584	There are only items from M1 at this level. It is appreciated that the students possess historical knowledge which they use to contextualize and better understand the facts and questions they face. Knowledge of the chronology and circumstances of the post Spanish Civil War allow them to consider as reasons for emigration the archaic nature of traditional agriculture (1/M1_1) and the reprisals against the losing side in the war (5/M1_4). They recognize the boost to the economy afforded by Spain's entry into the EU and the adoption of the euro (14/M1_9.2), and they partially explain the changes in Spanish society from 1960 to the present day (21.1/M1_14.1).
5	>584	Not enough data. Multiple choice item 20/M1_13 lets us think that some students can link certain women' mentality and role in society to the material conditions of life during the post War period in Spain. Similarly, the open response item 21.2/M1_14.2 allows us to say initially that the students can, with the aid of the input documents, briefly but correctly explain the changes in Spanish society since 1960, with reference to the predominant economic, political and educational transformations.

Table 9. Levels of performance associated with the competence 'explaining historically'

Conclusion

The statistical analysis of our test confirms its validity as a tool for assessing Secondary students' acquisition of what we have called historical competences, based on historical thinking. This model of test could be considered a feasible way for history assessments in Spain to abandon rote learning-based questions as their main instrument. The model would have two major traits: first, the assessment of substantive or content knowledge should preferably be functional or applied to cases or facts not previously studied; and, second, the assessment of skills related to historical thinking should occupy a key position in the test. The adoption of this model of assessing History, be it on nowadays large scale sample diagnostic assessments in Spain, or in classroom test and exercises, may have a powerful impact on the curriculum and on the teaching of History, since it will spark teaching approaches more innovative and attractive to the students than traditional rote learning teaching. These approaches require problem-based learning, while at the same time emphasize the value of history in understanding and thinking about our present. Finally, in relation to our second research question, the design of the test and statistical analysis of responses seem to afford a basis on which to build a scale of progress in the acquisition of historical thinking. As international assessment studies exhibit, this kind of analyses allows to build a scale of progress based on the varied difficulty of items and responses. The greater number of items and responses we get, the richer and more comprehensive the scale of progress will be. In accordance, the adoption of this model of History assessments in Spanish sample diagnostic test, could boost research on students' historical thinking. They would provide a considerable amount of responses and reflections on historical problems and issues with which enrich students' acquisition of historical thinking and, thus, the education professional knowledge.

References

Arias, L., Egea, A., Sánchez, R., Domínguez, J., García, F. J., & Miralles, P. (2018). ¿Historia olvidada o historia no enseñada? El alumnado de Secundaria español y su desconocimiento sobre la Guerra Civil. *Revista Complutense de Educación, 30*(2), 461-478. http://dx.doi.org/10.5209/RCED.57625

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

Domínguez, J., Arias, L., Sánchez, R., Egea, A., & García, F. J. (2016). Cómo evaluar el pensamiento histórico en la ESO: primeros resultados de una prueba piloto. In R. López-Facal (ed.), *Ciencias sociales, educación y futuro. Investigaciones en didáctica de las ciencias sociales* (pp. 176-187). Santiago de Compostela: USC.

Duquette, C. (2015). Relating historical consciousness to historical thinking through assessment. In K. Ercikan, & P. Seixas (Eds.), *New directions in assessing historical thinking* (pp. 51-63). New York & London: Routledge.

Eliasson, P., Alvén, F., Yngvéus, C. A., & Rosenhud, D. (2015). Historical consciousness and historical thinking reflected in large-scales assessment in Sweden. In *New directions in assessing historical thinking* (pp. 171-182). New York & London: Routledge.

Embretson, S. E., & Reise, S. P. (2000). *Item response theory for psychologists.* Mahwah, NJ: Lawrence Erlbaum Associates Publishers.

González, N., Pagès, J., & Santisteban, A. (2011). ¿Cómo evaluar el pensamiento histórico del alumnado? In P. Miralles, S. Molina, & A. Santisteban (Eds.), *La evaluación en el proceso de enseñanza y aprendizaje de las Ciencias Sociales* (vol. 1, pp. 221-232). Murcia: AUPDCS.

Körber, A., & Meyer-Hamme, J. (2015): Historical thinking competencies and their measurement: Challenges and approaches. In K. Ercikan & P. Seixas, *New directions in assessing historical thinking* (pp. 89-101). New York & London: Routledge

Lee, P. J. (1983). History teaching and philosophy of history. *History and Theory, 22*(4), 19-49. https://doi.org/10.2307/2505214

Lee, P. J. (2005). Putting principles into practice: Understanding history. In M. Donovan, M. Suzanne, & J. D. Bransford (Eds.), *How students learn. History in the classroom* (pp. 31-77). Washington DC: National Academies Press.

Lee, P. J., & Shemilt, D. (2003). A scaffold, not a cage. Progression and progression models in history. *Teaching History, 113*, 13-23.

Lévesque, S. (2008). *Thinking historically educating students for the twenty-first century*. Toronto: University Toronto Press.

OECD (2017). *PISA 2015 Assessment and analytical framework: Science, reading, mathematic, financial literacy and collaborative problem solving*. Paris: OECD. https://dx.doi.org/10.1787/9789264281820-3-en

Sáiz, J., & Fuster, C. (2014). Memorizar historia sin aprender pensamiento histórico: las PAU de Historia de España. *Revista Investigación en la Escuela, 84*, 47-57. Retrieved from http://hdl.handle.net/11441/59755

Seixas, P. (1996). Conceptualizing growth in historical understanding. In D. R. Olson & N. Torrance (Eds.), *The handbook of education and human development: new models of learning, teaching and schooling* (pp. 765-783). Malden: Blackwell.

Seixas, P., & Peck, C. (2004). Teaching historical thinking. In A. Sears & I. Wright (Eds.), *Challenges and prospects for Canadian social studies* (pp. 109-117). Vancouver: Pacific Educational Press.

Seixas, P., & Morton, T. (2013). *The big six: historical thinking concepts*. Toronto: Nelson Education.

Seixas, P., Gibson, L., & Ercikan, K. (2015). A design process for assessing historical thinking: The case of a one-hour test. In K. Ercikan, & P. Seixas (Eds.), *New directions in assessing historical thinking* (pp. 102-116). New York & London: Routledge.

VanSledright, B. (2014). *Assessing historical thinking and understanding: Innovative designs for new standards*. New York: Routledge.

Wilschut, A. (2015). Testing frame of reference knowledge in national examinations: Report on an experiment in the Netherlands. In A. Chapman & A. Wilschut, *Joined-up history. New directions in history education research* (pp. 85-112). Charlotte: IAP.

Wineburg, S. S., & Steinberg, S. (2007). *Reading like a historian toolkit*. Orlando: Holt, Rinehart and Winston.

Endnotes

[1] https://eur-lex.europa.eu/LexUriServ/LexUriServ.do?uri=OJ:L:2006:394:0010:0018:EN:PDF

[2] The complete test in Spanish, as well as the translation into English together with tables detailing the historical competences and skills that each item tries to assess are available in https://www.um.es/dicso/es/cchh/.

[3] The specific items of the test can be consulted (both in English and in Spanish) in https://www.um.es/dicso/es/cchh/

Acknowledgements

This research was supported by Ministerio de Economía y Competitividad, Spain / FEDER within the frame of the project *La evaluación de las competencias y el desarrollo de capacidades cognitivas sobre Historia en Educación Secundaria Obligatoria* [EDU2015-65621-C3-2-R].

About the Authors

Jesús Domínguez-Castillo holds a B.A. in History and Geography (Universidad de Zaragoza, 1974), a M.A. in History in Education (UCL, Institute of Education, 1984), and a PhD in History (Universidad Autónoma de Barcelona-UAB, 1993). He was head of department of Geography and History, and Director of Spanish State Secondary Schools. Advisor to the Ministry of Education (1988-94), he coordinated the LOGSE 12-16 curriculum of Social Sciences, Geography and History. He also was National Research Coordinator of IEA PIRLS and TIMSS (2008-2012). In his last professional years, he has been associate researcher and MA teacher at the University of Nebrija, Madrid. His major professional interests are History teaching and assessment of students' historical thinking. One of his main contributions has been *Pensamiento histórico y evaluación de competencias*, published in 2015 in Barcelona (Ed. Graó) (Its appendix test is available in English at https://www.um.es/dicso/es/cchh/).

Laura Arias-Ferrer is Associate Professor [Profesora Contratada Doctora] in the Department of Mathematical and Social Sciences Teaching-University of Murcia (Spain). Her work is focused on the analysis of the teaching practice and the training of Early Childhood and Primary Education teachers, as well as on the development and evaluation of strategies and resources for the teaching of history at different educational levels, with special attention to those methodologies focused on the development of historical thinking skills through the use of cultural heritage. These lines of work have led her to carry out stays at the universities of Zaragoza, Barcelona, Kentucky and London, whose fruits are the various works published regarding the mentioned topics. Her expertise has allowed her to be actively involved in various dissemination projects through the dissemination initiative *LATE. Laboratorio Temporal* (https://www.um.es/late/), coordinated by Alejandro Egea and Laura Arias.

Raquel Sánchez-Ibáñez is currently Associate Professor [Profesora Contratada Doctora] in the Department of Mathematical and Social Sciences Teaching at the University of Murcia (Spain), where she held different posts since 2009. She has spent abroad as visiting scholar a total of 28 months in prestigious centres such as the School of History (University of East Anglia), Instituto de Ciencias Sociais de Lisboa (Portugal), Università degli studi di Bari (Italy) and Centro Internazionale di Didattica della Storia e del Patrimonio (Bologna, Italy). Her research focuses on history teaching, social science teacher training and textbook analysis. She has participated in 11 funded research projects on the history and didactics of the social sciences and in 10 research projects on educational innovation, of which she has conducted 6. Finally, in the last ten years she has published 20 articles in high impact journals.

Alejandro Egea-Vivancos is Associate Professor (tenured) [Profesor Titular] in the Department of Mathematical and Social Sciences Teaching-University of Murcia (Spain). Historian, archaeologist and former teacher in Secondary Education, he holds a MA in Ancient History and was awarded his PhD in 2003. His work currently focuses on innovation in teaching practice by introducing archaeology and heritage, the development of historical thinking in early years & secondary education. These lines of work have led him to carry out stays at the universities of Barcelona, Kentucky and London. His expertise as Historian, History teaching researcher and his former teaching experience has allowed him to be actively involved in various dissemination projects through the dissemination initiative *LATE. Laboratorio Temporal* (www.um.es/late/), coordinated by Laura Arias and himself. He is the main editor of *Panta Rei. Journal of History and History Teaching*.

Francisco Javier García-Crespo was teacher of Secondary Education between 1993 and 2009 and Technical Advisor to the National Institute of Educational Evaluation between 2009 and 2011. He is Head of the Division of Management and Data Analysis (INEE) since 2011 and part-time professor in the Department of Statistics and Research Operations, Faculty of Mathematics, Complutense University of Madrid (Spain), where he teaches Statistics, Probability and Research operations. Among other positions, he is National Research Coordinator TIMSS2015, Data Manager TIMSS2011, PIRLS2011 and PIRLS2016, Activity Coordinator for the Large scale data analysis training for teachers and researchers. He has also participated as speaker in the presentation activities for international studies and he is author of 7th and 8th grade textbooks. He has also participated in the development of various national reports of the General Diagnostic Assessments (primary and secondary) and international (TIMSS-PIRLS 2011, ICCS2009, PISA 2012 and TALIS 2013).

Pedro Miralles-Martínez has hold different teaching position for 38 years in various educational levels: Primary Education; Geography and History teacher in Secondary Education; and professor at university. He worked in permanent teachers' training in Teachers and Resources Centres. He is the Principal Investigator to the research group 'Didactics of Social Sciences'. He has directed or participated in more than twenty research and innovation projects on didactics of social sciences in Education. He is the author of more than a hundred papers and more than a hundred conferences and communications in scientific meetings. It is remarkable the publication in JCR and SCOPUS journals.