Arteaga-Martínez, B., Macías-Sánchez, J., Pla-Castells, M., & Ramírez-García, M. (2021). Design and validation of an observation instrument for mathematics classes in secondary education: Expert panel and Delphi process. *RELIEVE, 27*(2), art. 3. http://doi.org/10.30827/relieve.v27i2.21812

Revista ELectrónica de Investigación y EValuación Educativa

RELIEVE

ISSN: 1134-4032

e-Journal of Educational Research, Assessment and Evaluation

Design and validation of an observation instrument for mathematics classes in secondary education: Expert panel and Delphi process

Diseño y validación de un instrumento para observación de clases de matemáticas en Educación Secundaria: grupo nominal y método Delphi

Arteaga-Martínez, B. (1), **Macías-Sánchez, J.** (2), **Pla-Castells, M.** (3), **& Ramírez-García, M.** (4)

(1) Spanish National University of Distance Education, UNED (Spain), (2) Complutense University of Madrid (Spain), (3) University of Valencia (Spain), (4) La Salle University Centre (Spain).

Abstract

Observation and interpretation processes are rarely used in teaching practice as learning tools in Spain. In order to encourage their use, it is important to have instruments that facilitate analysis in practice and that gather information about the particular characteristics of the educational context in which they are used. The present study presents the process of design, construction and validation of POEMat.ES, an observation and interpretation instrument for use in relation to the teaching of secondary education mathematics classes. This process consisted of two phases: 1) literature review and design using an expert panel made up of 24 specialists; 2) construction and validation via the Delphi method conducted by 15 experts over two rounds. The final version of the tool was organized around three dimensions: mathematical content, didactics of mathematical content and classroom management. It comprised 17 indicators, each of which was classified according to four levels. The methodology applied conformed to consensus, with this being considered adequate for the selection and validation of the indicators that frame the observation of teaching practice. I can be concluded that POEMat.ES is a useful tool for systematizing the observation of teaching practice in order to understand and accurately interpret the process of teaching mathematics in secondary education in the Spanish context.

Keywords: Delphi method, expert panel, mathematics teaching, secondary teaching, observation tool.

Resumen

El uso de los procesos de observación e interpretación de la práctica docente como herramienta de aprendizaje para los docentes es infrecuente en España. Para promover su utilización, se considera la importancia de contar con instrumentos que faciliten este análisis de la práctica, a la vez que recogen las características particulares del contexto educativo donde se utilizan. Este estudio presenta el proceso de diseño, construcción y validación de POEMat.ES, un instrumento de observación e interpretación de la práctica docente, en las clases de matemáticas de la etapa de Educación Secundaria. Este proceso consta de dos fases: 1) revisión de la literatura y diseño, utilizando un grupo nominal con 24 especialistas; 2) construcción y validación mediante un método Delphi en dos rondas, con 15 expertos. La versión final del instrumento se organiza en torno a tres dimensiones: contenido matemático, didáctica del contenido matemático y gestión del aula, con un total de 17 indicadores, cada uno de los cuáles se clasifica en cuatro niveles. Las metodologías de consenso utilizadas se consideran adecuadas para la elección y validación de los indicadores que enmarcan la observación de la práctica docente. Se concluye que POEMat.ES es una herramienta útil para sistematizar la observación de la práctica docente y así comprender e interpretar con precisión el proceso de enseñanza de las matemáticas en el aula de matemáticas de Secundaria según la realidad española.

Palabras clave: método Delphi, grupo nominal, enseñanza de las matemáticas, educación secundaria, instrumento de observación.

Received/*Recibido*	Jul 20, 2021	Approved /*Aprobado*	Nov 29, 2021	Published/*Publicado*	Dec 22, 2021

Corresponding author / Autor de contacto: Blanca Arteaga Martínez. Spanish National University of Distance Education, Faculty of Education. C/ Juan del Rosal 14, 28040, Madrid (Spain). correo-e: blanca.arteaga@edu.uned.es

Research in mathematics teaching has used classroom observations as a main approach to analysis. This makes it a powerful tool (Bostic et al., 2021) for professional development, given the enormous impact of teaching performance on student performance (Weber et al., 2018).

Thus, the present work sets out by considering the types of observational tools currently being used in mathematics classrooms in secondary education.

First, an initial search of specialized literature on the didactics of mathematics was conducted to identify observation instruments. The first conclusion from this review pointed to a scarcity of specific guidelines for mathematics teaching, with general instruments developed within different subjects typically being applied (Schlesinger & Jentsch, 2016). This approach likely detracts from the accuracy of observation. In this sense, observational tools were identified which covered aspects such as classroom management, specific attention and cognitive activation, with the latter two specifically targeting students. These three dimensions have been considered in other studies on mathematics teaching, which revealed deficiencies in data collection regarding the mastery of the subject (Praetorius & Charalambous, 2018). Observation tools to have been analyzed in depth include the mathematical quality of instruction ([MQI] Hill et al., 2008), instructional quality assessment ([IQA] Matsumura et al., 2008) and the *pauta de observación de clases de matemáticas impartidas por profesores principiantes PROMATE* [guidelines for observing junior teachers' mathematics classes] (Barriendos et al., 2018).

Other examined instruments to have also been proven useful for making an initial approach to the observation foci that could exist in a mathematics classroom include the reform-oriented teaching observation protocol ([RTOP] Piburn & Sawada, 2000), mathematics scan ([MScan] Walkowiak et al., 2014), elementary mathematics classroom observation form (Thompson & Davis, 2014) and international system of teacher observation and feedback ([ISTOF] Teddlie et al., 2006).

Following this initial analysis of observational tools, it was considered whether there were cultural elements that influenced mathematics teaching models. The review again showed that little research existed (Thomas & Berry III, 2019) that provided comparative data in transcultural contexts from the perspective of teaching, perhaps due to the "sense of the universality of the mathematical concepts" (Orton, 2003, p. 105). However, from a mathematics teaching perspective, different studies have examined cultural differences, for example, in student beliefs (Diego-Mantecón & Córdoba-Gómez, 2019) or problem solving (Eccius-Wellmann et al., 2017). Conversely, scientific literature on the didactics of mathematics tend to produce outcomes that indicate transnational differences in both didactics and learning (Openshaw & Walshaw, 2019).

Thus, as a general aim, the present study sought to develop an observation schematic adapted to the reality of secondary education classrooms in Spain in order to facilitate the examination of practice from the perspective of observed teacher actions through the recording of classroom sessions. The present article describes the construction and validation of this schematic, according to two stages which pertain to the following specific objectives:

1. Bring together a group of specialists to draft preliminary indicators using expert panel technique (NGT).
2. Use the Delphi method to analyze and validate the selected indicators and produce a definitive guideline.

Method

Methodology applied to instrument design

Previously conducted literature examining the application of observation instruments in mathematics classrooms was reviewed in order to form the basis of a preliminary schematic. This corresponded to three fundamental dimensions, which pertained to the entire spectrum of analysis of teacher performance during practice: mathematical knowledge, mathematical didactic knowledge and classroom management (Aguilar-González et al., 2018; Garzón, 2017; Van Zoest et al., 2021).

A coordinating team was set up to select a group of specialists who would work together in line with the NGT. The aim of this was to identify the indicators associated with each of these dimensions.

This group, selected according to a transnational and cross-cultural perspective, worked together for four weeks, taking advantage of a pre-arranged meet at an American university. The group consisted of mathematics lecturers from different universities, with all of them having taught on mathematics teacher education courses prior to exercising at a university. Participating lecturers came from different countries: Spain (n=17), the USA (n=3), Finland (n=3) and Israel (n=1). Structured face-to-face meetings were held, with the intention that individual ideas would be presented and eventually lead to group consensus.

The aim of this stage was to increase creative productivity of the group, inducing joint decisions from the stimulation of critical ideas through structured meetings that allowed all participants to be heard (McMillan et al., 2016) and to listen to others. This approach involves all agents in a social phenomenon which leads to the collective resolution of a complex problem (Moore, 1987). The four phases of the expert panel were based on a previously established structure which was communicated to participants at the outset.

The first phase of the elaboration of individual answers (silent generation) was based on questions around the three dimensions. Each dimension was addressed one by one and in writing in order to ensure that the ideas of each of the participants were captured. In the second phase, the coordinating team directly requested the input of each of the members face to face in order to generate new ideas. The coordinating team compiled all ideas by video-recording the sessions. These ideas were then presented digitally as a summary at the beginning of each session. The third phase, focused on clarification, sought to grouping together similar ideas. To this end, the instrument sections that had already been developed were used to conduct an observation and analysis of video fragments recorded in different secondary school classrooms in Spain. These sessions were conducted so that questions and doubts could be raised in order to make appropriate modifications to the indicators. In the final phase (voting), the panel quantitatively assessed recorded observations using the video fragments tool. This phase was conducted individually, digitally and anonymously.

Following this process, the observation instrument retained its initial three-dimensional structure, with the first two comprising seven indicators each and the third dimension comprising four indicators. Each indicator was to be rated with a numerical score that ranged from 0 to 3. In some cases, this does not conform to an ordinal scale but to a ranking.

Dimension 1, “Mathematical Content”, aims to capture the way in which the teacher articulates content intrinsic to mathematics through representations, mathematical flexibility, definitions and reasoning. Its first three indicators are based on the works of Duval (1993) into the use, treatment (transformations of mathematical content of the same order) and conversion of representations. The fourth assesses the depth with which the teacher defines and uses definitions in the classroom. The fifth focuses on teachers’ reasoning processes. The sixth, mathematical flexibility, aims to observe whether the teacher uses and promotes the ability to generate different strategies to complete a task, ideally comparing them explicitly and reflecting on their characteristics (Schneider et al., 2011; Liu et al., 2018). A final indicator was included which shows whether the teacher made a mistake and, if so, provides a description.

Dimension 2, “Didactics of Mathematical Content”, seeks to study the specific aspects of teaching-learning processes applied to mathematics. The indicator “use of materials” aims to observe whether the teacher uses any objects, materials, or technical means to help students understand mathematical content and encourage them to practice mathematics (Maz-Machado et al., 2019). In response to the nature of required tasks, teachers tend to operate within a double dichotomy determined by the degree of balance between what is requested and what is ultimately provided (Molina & Samper, 2019). This is met by demand for accessibility, with this balance depending on whether students possess the mathematical knowledge necessary to solve problems or whether such knowledge must be created. The indicator sequencing and connections focuses on the chain of tasks and the way in which they are connected to other mathematical content in the arguments made by the teacher. The contextualization of mathematical content aims to describe the use of "extra-mathematical" contexts that endow mathematical notions with meaning and help students acquire conceptual knowledge on a given set of mathematical content. This includes the set of situations and concepts that form these notions and improve its understanding and functionality (Kaiser, 2020; Vergnaud, 2013). The indicator pertaining to the ending of responsibility for mathematical activity seeks to capture the degree of student autonomy promoted by teachers through the discovery and assessment of strategies. This indicator is based on the theory of didactic situations (Brousseau, 2007). The indicator mathematical language describes whether teachers use formal language which may combine words, symbols and figures with mathematical meaning (i.e., determined by very precise rules and relationships based on logical deduction), whilst also adapting language to the students' educational level (Planas et al., 2018). The indicator pertaining to the didactic-mathematical exploitation of student input describes teachers' use of student mathematical content input. It is difficult to predict when such situations will arise (Rowland et al., 2009) but their emergence can be used to measure the quality of classroom instruction as they demonstrate student opportunities to participate in mathematical discussions (Boston & Candela, 2018).

Dimension 3, "Classroom Management", aims to observe aspects of classroom practice that intervene in the teaching-learning process without referring to specific aspects of the mathematical discipline. The indicator density measures the ratio of time spent on teaching-learning mathematics in relation to overall time without assessing quality. MQI's previous work shows that the amount of time spent on mathematical activity and classroom management plays an important role on the quality of teaching (Hill et al., 2008). The indicator resource management is intended to describe the clarity, efficiency and effectiveness of the use of material resources. The use of written material describes whether teachers use written materials. In Spain this practice is common but, in other countries, teachers do not always depend as heavily on textbooks (van den Ham & Heinze, 2018). Finally, disruptive behavior management aims to observe the effectiveness of teacher management when faced with situations caused by students that do not favor session development (Rodríguez & Ruiz, 2019).

Instrument assessment: The Delphi method

The Delphi method was selected to conduct the second phase of work which was aimed towards content assessment, (Martino, 1999). The aim of this phase was to ensure the inclusion of all relevant elements and critical information which may be essential for explaining any situation likely to be observed at a given time (Okoli & Pawlowski, 2004). This methodology facilitates group thinking and prevents the opinions of dominant panel members from being overrepresented.

Questionnaires that assessed dimensions and indicators were used as data collection instruments, alonside controlled comments reflecting selected opinions. Data collection was planned to follow a flexible, albeit limited, timeframe, which facilitated the participation of a geographically dispersed group with different timetables (López-Gómez, 2018).

Selection and shaping of the expert panel

The work began with the formation of the group of experts, prioritizing the representativeness of participants over their number. Criterion sampling was used for this selection process (Patton, 1990). The expert panel was selected following the definition of criteria that address the balance between current knowledge and perceptions of mathematics teaching. The initial instrument dimensions established as a fundamental criterion that experts be both mathematics specialists and teachers actively working as educators.

In order to meet these criteria and recruit specialists in the didactics of mathematics, 20 experts from different Spanish universities were invited to participate. During this process, three questions were posed to assess their predictive capacity and commitment to the process: years of professional experience, description of their teaching and research work, and self-assessment in relation to the project objective.

To assess suitability, expert competence indices (K) were calculated. This index is made up of two coefficients: a knowledge index (Kc) and an argumentation index (Ka). This index enables determination of the suitability of expert inclusion in the panel (Llorente, 2013). An acceptable K was greater than .8 (Cabero & Barroso, 2013).

In order to obtain this index, the coordinating research team, at its discretion, weighted different argument types. Assigned weights enabled a quantitative assessment of expert self-report responses to be conducted. In order to calculate Kc, each participant was asked to rate, from 0 to 10, the following question: On a rising scale from 0 to 10, how would you rate your level of training (through the knowledge acquired both from your training and professional experience) for addressing the proposed research project? Kc was calculated by multiplying this score by .1. In order to calculate Ka, participant responses to the following question were collected: To what degree (greatly, moderately, little) will each of the following aspects related to your professional background be important to your arguments when participating in the research project? These aspects were entered into a scoring crib sheet, which had been organized a priori according to the expected structure of the observation instrument (Table 1). After assigning the response selection with its corresponding weight, a summary score is produced.

Table 1. Summary scores for calculation of the argumentation index

Subject	Greatly	Moderately	Little
1. Mathematical content (on the topics addressed in the stage considered)	0.35	0.28	0.13
2. Didactic-mathematical content (on mathematics teaching and learning)	0.30	0.24	0.12
3. General pedagogical content	0.05	0.04	0.025
4. Mathematics teachers' initial or continuing education	0.10	0.08	0.075
5. Descriptive observation of classroom situations (own or other teachers')	0.10	0.08	0.075
6. Interpretative observation of classroom situations (own or other teachers')	0.10	0.08	0.075
Total	**1**	**0.80**	**0.50**

There are no clear rules on the optimal number of experts required to make up these panels. Previous research indicates that there must be a minimum number of 10 experts (Parenté & Anderson, 1987). Landeta (1999) proposes between 7 and 30, and a review by Ludwig (1997) indicates that most studies use between 15 and 20. Thus, a total of 15 experts was decided to be adequate for the present study (Clayton, 1997). Their motivation to commit to the entire duration of the validation process was also taken into account (Buck et al., 1993).

Selected experts were all exercising as educators at one of seven Spanish universities. The sample was composed of six men and nine women, 86.67% had PhDs and 93.33% had a degree in mathematics.

The group was highly heterogeneous. All had university teaching experience (M=17.86 years) and research experience in the field of the didactics of mathematics (M=9.36 years).

In addition, seven experts had teaching experience in secondary education.

Two-round assessment in line with the Delphi method

Members of the coordinating team were responsible for collecting information from experts in an individual and anonymous way through an iterative process. This process involved "returning the proposed set of indicators to experts for review and agreement" (Cabero & Llorente, 2013, p.17) and then employing measures of central tendency to quantitative data pertaining to indicators. Communication between the coordinating team and each of the panelists was done via email.

This method ensured that each of the experts was given space to reconsider their initial response following discussion of the contributions made by other panel members. This means that "[the] iterative process advances at the same time that consensus or agreement values are obtained in relation to the statements raised or, if this does not occur, positions in which manifest discrepancies are observed are consolidated" (López-Gómez, 2018, p. 26).

Two rounds were planned meaning that the process could be considered as a "modified Delphi" method, which seeks to reach an optimal consensus without causing fatigue in participants (Cabero, 2014) and assumes that experts work on a topic that has not been generated by themselves and on which agreements must be reached (Cabero & Infante, 2014). Two rounds is an appropriate number to ensure convergence (Linstone & Turoff, 1975).

Both consensus and disagreement can be rated in this type of process with respect to each of the elements of the instrument as all expert contributions are considered to be important (Gordon, 1994). Expert judgments are enriched by the process and the information they receive, meaning it can be assumed that later contributions will be richer than early ones, eventually resulting in an optimal group response (López-Gómez, 2018). Experts were geographically separated and so worked individually and anonymously. The first round began with the individual application of the instrument to a sample of video fragments produced from sessions pertaining to secondary school classrooms (each researcher viewed 29 fragments of between 10-15 minutes) which had been selected by the coordinating team. Experts provided qualitative and quantitative information in relation to each of the instrument's indicators. The coordinating team sent a document to each panel member, in which the video fragments they were to view were presented. Fragments were selected in a way that ensure that each expert would view fragments with the same characteristics in terms of topic, academic year, class time (beginning, middle or end) and teacher. Application of this tool provided the necessary stimulus for panel members to generate ideas and initiate the decision-making process. In this first round, each expert was asked to feedback on their review of the existing literature and use of the instrument. An assessment questionnaire was sent to each expert to be completed within three weeks at the latest. Each expert was asked to rate on a 10-point scale the importance of each of the instrument's indicators for responding to the research issue. An open-ended question was also included to allow experts to provide comments or justify their ratings.

The second round was conducted on the basis of the outcomes produced during the first round and a report sent to all experts regarding potentially pertinent modifications. This document detailed the changes suggested following classification of the individual responses given anonymously in round 1. Collected data were based on open questions and conformed to the structure of the instrument. This approach ensured that the coordinating team did not omit any relevant issue. "Upon commencing the second round, experts are asked to consider the location of the panel's measure of central tendency and are allowed to review their initial answers if they so wish" (Novakowski & Wellar, 2008, p. 1486).

The interval between rounds was three months. The final outcomes of this round gave rise to consensus and led to the final instrument.

Results

In the presentation of the final outcomes produced from the expert panel, a graphical presentation of the quantitative outcomes from both rounds was opted for, alongside the textual description of the instrument. An explanatory document was also drafted which detailed the design of each of the two Delphi rounds and the contributions made.

First round outcomes (Table 3) were similar with regards to the first two dimensions, with values pertaining to the third dimension being slightly lower and also showing greater dispersion between indicators. Two dimensions (<15%) showed little, while one (15-30%) had moderate dispersion (Martins & Theóphilo, 2007).

Following examination of round 1 data, some experts were requested to justify their low scores for some of the indicators in cases where dispersion between values was high compared with the prevailing opinion of the group (Clayton, 1997). The outcome of this round made it easier for the coordination team to develop a new version of the instrument. In addition to modifications for each indicator, structural change suggestions were also gathered, such as the need to define the terms “topic” and “task”, which were included in the introduction of the instrument. The entire process is presented in Table 2, detailing the intervention according to indicator and the round in which the contribution was made. It should be mentioned that one indicator was discarded following the first round. Each expert was provided with the group mean, a graphical representation of the frequency distribution and individual responses provided during the first round, followed by a detailed overview of general findings pertaining to each of the dimensions.

Table 2. Summary of instrument modification suggestions emerging from both rounds of the Delphi method

Final instrument indicator	First round changes	Second round changes
Use of representations (1.1)	Levels are described and the registers pertinent to each are clarified. The natural language (NL) register is differentiated. Examples given to assist in the identification of ratings are modified.	It is specified that different registers should not be used for the same content.
Treatment of representations	Doubts were expressed about the need to maintain this aspect due to the information contributed and the degree of specialization reviewers must have on representation registers to be able to apply the indicator. Discarded.	
Conversion of representations (1.2)	Level descriptions are simplified based on the number of registers required to make conversions for the same content, indicating the meaning of the conversion and the relationship with NL. Text outlining the brief explanation is improved and examples modified.	NL representation discarded.
Definitions (1.3)	The process of defining teachers is condensed. The brief explanation is modified to clarify the aim of indicators. Rating scores and examples are modified.	The rigor of teacher-provided definition will now be assessed in relation to whether they address characteristics of mathematical definitions. The brief explanation, level descriptions and examples are modified, aspects such as the use of definitions to support arguments is considered in 1.4.
Argumentation (1.4)	The name is changed from reasoning to argumentation. The aim of the indicator is clarified and explained in the brief explanation. Level descriptions and examples are modified.	Definitions are included within the brief explanation.
Mathematical flexibility (1.5)	The brief explanation and rating scores are modified to clarify the difficulties noted in rating. Some examples are modified to facilitate scoring.	

(Table 2. Continued)

Final instrument indicator	**First round changes**	**Second round changes**
Connections (1.6)	The indicator previously pertaining to dimension 2 named "sequencing and connections" is changed to "connections", and now corresponds to the first dimension in which it describes inter-relationships between mathematical themes. The brief explanation and levels are modified, eliminating everything related to sequencing and redrafting everything in reference to connections within mathematics. New examples are provided.	Assessment categories provided to collect information on the way in which teachers connect topics and mathematical phenomena are described more precisely. The brief explanation, levels and examples are modified.
Teacher mathematical errors (1.7)	This indicator is thoroughly reformulated, moving from requiring a dichotomous yes/no response to gathering teachers' mathematical errors when responding to student inputs. The different types of errors to be observed are presented. Levels are replaced with examples.	Six levels are included, where 0 describes the absence of errors and 1-5 pertain to types of mathematical errors often made by teachers when responding to students.
Use of materials (2.1)	Levels are modified to focus on the way in which teachers put materials into practice.	Slight modifications are made to the brief explanation to distinguish the use of materials from other written material included in dimension 3.
Nature of proposed tasks (2.2.)	The notions of "accessible" or "not-accessible", and "closed" and "open" are clarified.	The word "proposed" is included in the name.
Contextualization of mathematical content (2.3)	This indicator is modified for observation of only the contextualization of mathematical content outside of the area. No sequencing is observed.	Level 0 is reimagined to specify that content appearing in the fragment should not be contextualized at all. Examples are written to place focus on the teacher.
Transfer of responsibility for mathematical activity (2.4)	Level descriptions are modified to identify whether responsibility is transferred and the conditions required.	Examples are added and improved according to the new level descriptions.
Discourse adequacy (2.5)	The name is changed from "mathematical language" to "discourse adequacy". Focus is diverted to the adaptation of language to students' educational level. The brief explanation is modified, incorporating level descriptions and examples.	
Exploration of student input (2.6)	The brief explanation is expanded and examples are slightly modified to describe aspects such as quality, intensity and other common aspects pertaining to level descriptions. Level 0 is modified to reflect the absence of input or contributions lacking in mathematical content.	The brief explanation is expanded and examples are modified to indicate aspects such as quality and intensity.
Density (3.1)	Examples are modified to help distinguish the difference between rating scores 1-3.	Levels are defined according to time intervals devoted to mathematical content without going into intensity and quality, incorporating student autonomous work. The brief explanation, level descriptions and examples are modified.
Use of presentation resources (3.2)	The name is changed from "resource management" to "use of presentation resources" on order to focus on the clarity of presentation.	The level descriptions and brief explanation are modified to clarify the approach to be taken and improve the intelligibility of presentations. Example images are modified.
Use of written material (3.3)		The explanation is modified to indicate that only what can be observed in the use of written material is to be rated and intensity should not be considered. Level descriptions are modified to cover all possible motives of written material. Examples are modified.
Disruptive behavior management (3.4)		Disruptive behaviors that hinder teaching-learning processes in the classroom are identified. Teacher effectiveness in resolving conflict is assessed. The brief explanation, level descriptions and examples are modified to conform to the new approach.

Note. The complete guideline can be consulted at https://roderic.uv.es/handle/10550/78572

After making the pertinent modifications and continuing with the process, outcomes of the second round indicated a higher degree of consensus (Table 3). Average and the overall ratings pertaining to each dimension increased and dispersion reduced.

Table 3. Descriptive statistics for each dimension produced in the two rounds of the Delphi method

	First round			Second round		
Dimension	**M**	**s**	**CV**	**M**	**s**	**CV**
1. Mathematical content	9.181	1.017	11%	9.495	0.652	7%
2. Didactics of the mathematical content	9.167	0.974	11%	9.556	0.689	7%
3. Classroom management	8.400	1.330	16%	9.200	0.840	9%
Total	**8.992**	**1.129**		**9.447**	**0.724**	

As a part of the Delphi method, estimations are made to determine variation in the ratings given by each expert between the different rounds. Data is first presented in terms of the descriptive data pertaining to each of the indicators (Table 4).

Table 4. Descriptive statistics pertaining to each indicator produced in the two Delphi rounds

	First round						Second round					
Indicator	**M**	**s**	**CV**	**Mo**	**Mdn**	**IR**	**M**	**s**	**CV**	**Mo**	**Mdn**	**IR**
1.1.	9.5	0.6	0.1	10	10	1.0	9.7	0.5	0.1	10	10	0.5
1.2.	9.3	0.7	0.1	10	9	1.0	9.5	0.5	0.1	10	10	1.0
1.3.	8.9	1.1	0.1	10	9	2.0	9.3	0.7	0.1	9	9	1.0
1.4.	9.1	0.9	0.1	10	9	2.0	9.6	0.5	0.1	10	10	1.0
1.5.	9.3	1.1	0.1	10	10	1.0	9.6	0.6	0.1	10	10	1.0
1.6.	8.6	1.4	0.2	10	9	2.0	9.1	0.8	0.1	10	9	1.5
1.7.	9.4	0.9	0.1	10	10	1.0	9.6	0.7	0.1	10	10	0.5
2.1.	8.9	1.4	0.2	10	9	2.0	9.4	0.8	0.1	10	10	1.0
2.2.	9.1	0.8	0.1	9	9	1.5	9.7	0.6	0.1	10	10	0.5
2.3.	9.3	0.9	0.1	10	9	1.0	9.4	0.8	0.1	10	10	1.0
2.4.	9.3	0.8	0.1	10	9	1.0	9.6	0.6	0.1	10	10	1.0
2.5.	9.1	1.2	0.1	10	9	1.5	9.4	0.7	0.1	10	10	1.0
2.6.	9.5	0.6	0.1	10	10	1.0	9.9	0.4	0.0	10	10	0.0
3.1.	8.8	1.0	0.1	9	9	1.5	9.3	0.9	0.1	10	10	1.0
3.2.	8.3	1.0	0.1	8	8	1.0	9.1	0.7	0.1	9	9	1.0
3.3	7.9	2.0	0.3	8	8	1.5	8.9	1.1	0.1	9	9	2.0
3.4.	8.6	1.1	0.1	8	9	1.0	9.5	0.5	0.1	9	9	1.0

It can be noted that inter-rater agreement between rounds converges. The following comments can be made in relation to the data produced:

1. Interquartile range (IR). The calculation of IR is one of the tools used in the Delphi method (von der Gracht, 2012). In the present case, the IRs produced from data collected in the second round are lower than those produced in the first. A balance can be considered to have been reached, except in the case of indicator 3.3, which presented a greater IR in the second round. The relative

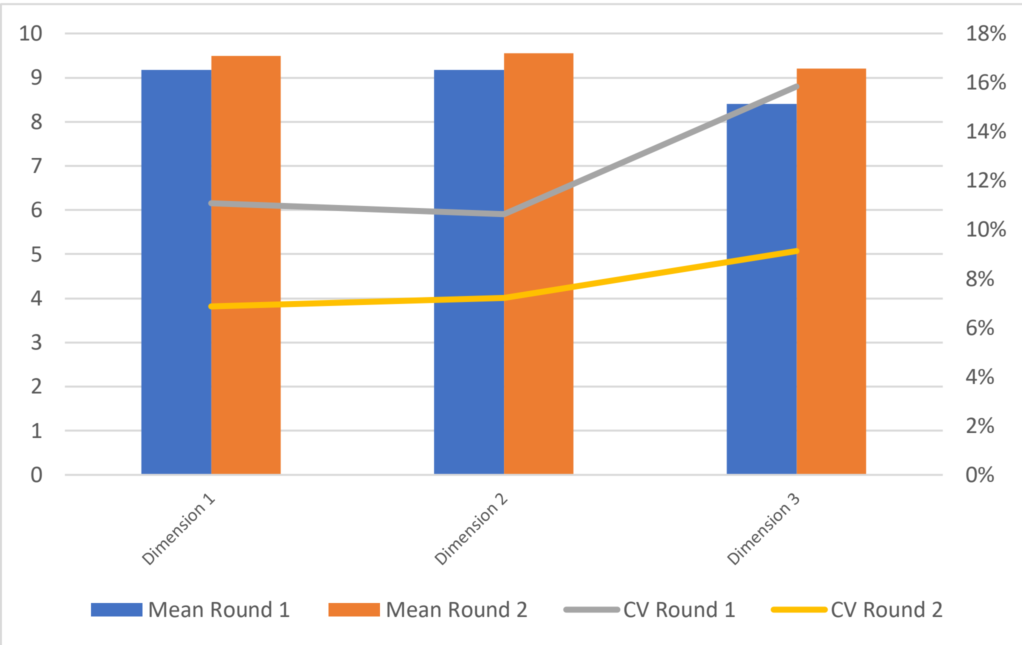
interquartile range (RIR) has been previously established as a single acceptance rating for summarizing all indicators (Landeta, 1999). This is defined as the ratio between the IR and the median. Using this statistic, it is possible to contextualize the IR in accordance with a central tendency value. Group stability is considered to be reached if variation in the RIR between rounds is less than .30 (Mengual et al., 2016). This condition was met for all indicators.

2. Coefficients of variation (CV), a measure without units and considered the best parameter for this type of study (Shah & Kalaian, 2009), were calculated and decreased in all cases showing that dispersion had converged between the scores given in the two rounds. Indicator 3.3 was, again, the indicator with the greatest dispersion.

These outcomes, in addition to the medians and modes produced for each of the indicators, point to optimal outcomes regarding inter-rater agreement and stability. This enabled the process to be concluded following this second round.

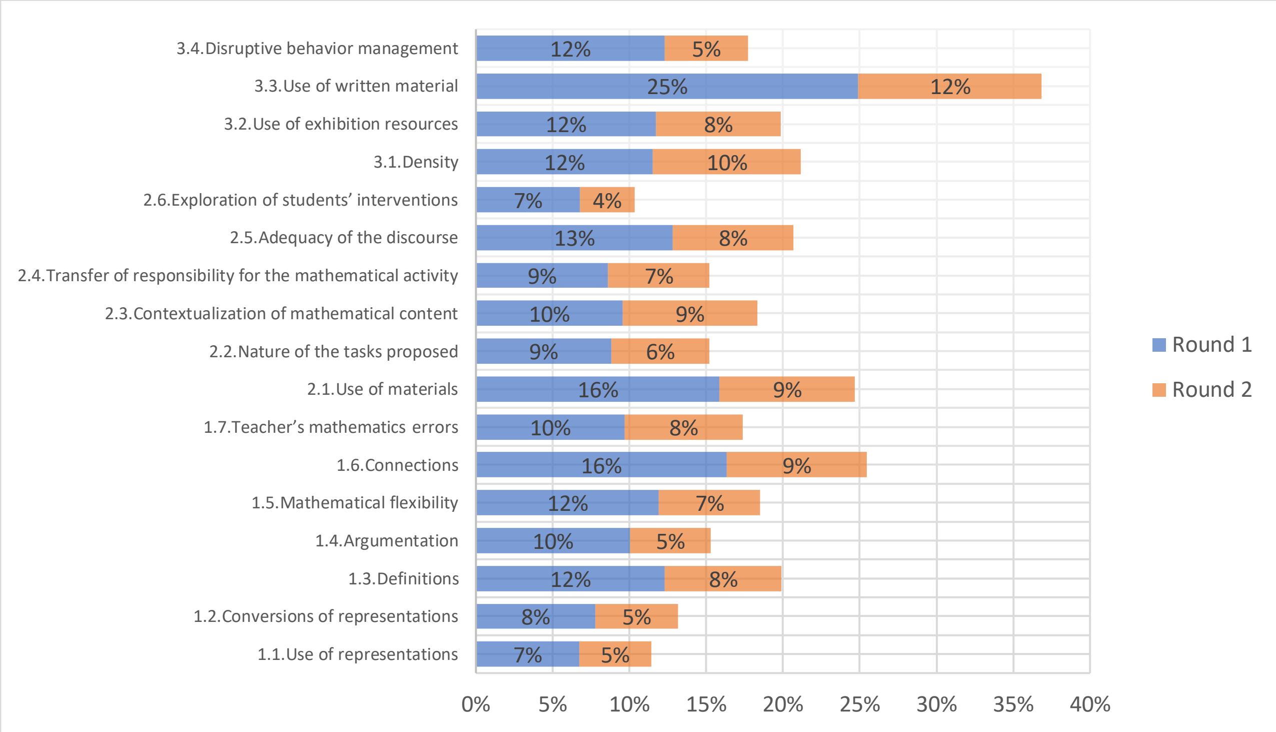
Overall, data (Figure 1) show an improvement on both accounts, with both an increase in the average ratings given to each of the indicators and reduced dispersion in these ratings. A CV>30% for any given indicator would indicate that the indicator must be discarded (Martínez, 1988). In all cases, CVs were less than 15%.

Figure 1. Means and coefficients of variation produced from each round according to dimension



For a more in-depth look at each of the indicators, important given that the aim of the present study was to establish agreement, variability between indicators was examined (Figure 2).

Figure 2. Dispersion between indicator ratings given in rounds 1 and 2



Conclusions

When considering the initial issues that gave rise to the development of the present work, it was noted that few observational tools were available for use in mathematics classrooms in secondary education. This being said, the issue at hand is an emerging research issue to which efforts are being dedicated within the research community, as can be seen in the special issue of the ZDM journal (Charalambous & Praetorius, 2018). The design of an instrument to facilitate the description of mathematics teachers' teaching practice is necessary for improving training programs whilst also serving as a reflection tool.

The construction of an instrument to guide analysis of teaching practice through the recording, analysis and interpretation of classroom sessions constitutes a necessary basis for initial teacher training and improvement from the point of view of didactics and the analysis of mathematical content.

With regards to the second proposed research issue and after having worked with specialists from different countries who have facilitated the configuration of the structure of POEMat.ES, it can be indicated that certain cultural elements influence teaching models in mathematics.

Taking the joint approach of considering the specific didactics and particularities underlying mathematics content, enables specialized analysis that is focused on teachers as conduits that align curricular content with students' learning possibilities, whilst also enabling a consideration of the specific Spanish context, which induces particular phenomena such as the use of textbook as standard elements of teaching practice (Dolores & Ibáñez, 2020; Monterubio & Ortega, 2009). The first of the specific stated objectives was achieved following review of observation instruments of mathematics classes, the establishment of a group of specialists and outcomes of the expert panel. This enabled an instrument to be developed which was adapted to the school and social environment in which it was going to be used.

The developed tool is adapted to the reality of classrooms in Spain and designed to collect key data via observations of practices in line with three dimensions: mathematical content, didactics of mathematical content and classroom management (whose management depends fundamentally on teaching action).

The methodological aspects of instrument development were adapted to the characteristics of the research group and the availability of resources at all times, being consistent with methodological guidelines guiding the achievement of the study objective. This was true both for the initial phase in which the expert group was established and the second phase in which the instrument was improved and validated using the Delphi method.

The instrument elaboration processing using the expert group resulted in 17 indicators distributed according to three dimensions. Each indicator is assessed according to a category score that ranges from 0 to 3. Following evaluation using the Delphi method, it was decided that these indicators would continue to be assessed according to numerical values, however, some were moved to correspond to different dimensions or details pertaining to their descriptions were changed, especially with regards to wording or the examples used to illustrate them and help the teacher place rank their observations according to one category or another.

The second of the stated objectives was achieved following application of the Delphi method. Concretely, evidence suggested that the elaborated instrument was valid and capable of assisting "researchers in mathematical education to empirically analyze classroom observations" (Bostic et al., 2021, p.9).

The limitations of the present work include the number of experts included in the expert panel. Although the sample could be considered sufficient for the overall evaluation process, a larger sample would have provided richer data. Nonetheless, it should be commented that the expert panel has remained active since its inception, added strength and reliability to the work carried out. The second limitation pertains to the video material used during the first Delphi round for subsequent assessment of the instrument. Experts all viewed the same fragments, which had been carefully selected to avoid bias. Nevertheless, given the complexities of human research it is possible that these fragments influenced responses in some small way.

The present study proposes an observational tool, POEMat.ES, for future use during different school years over the course of secondary education provision in Spain. This would be useful to compare with other outcomes and develop a repository of good practice for application in the mathematics classroom.

Funding

Real Colegio Complutense (RCC) at Harvard University

Acknowledgments

We would like to acknowledge the Real Colegio Complutense (RCC) at Harvard for the institutional and financial support provided for this research as part of the RCC SIMIS Study Group: https://rcc.harvard.edu/studying-and-improving-mathematics-instruction-secondary-schools-spain-simis

References

Aguilar-González, A., Muñoz-Catalán, M.C., Carrillo-Yáñez, J., & Rodríguez-Muñiz, L. J. (2018). ¿Cómo establecer relaciones entre conocimiento especializado y concepciones del profesorado de matemáticas? *PNA, 13*(1), 41-61. http://dx.doi.org/10.30827/pna.v13i1.7944

Barriendos, A., Berger, B., Domínguez, E., & Martínez, M. V. (2018). *Manual PROMATE. Pauta de observación de clases de matemáticas impartidas por profesores* *principiantes*. Centro de Investigación Avanzada en Educación de Chile.

Bostic, J., Lesseig, K., Sherman, M., & Boston, M. (2021). Classroom observation and mathematics education research. *Journal of Mathematics Teacher Education*, *24*, 5–31. https://doi.org/10.1007/s10857-019-09445-0

Boston, M.D., & Candela, A.G. (2018). The Instructional Quality Assessment as a tool for reflecting on instructional practice. *ZDM Mathematics Education, 50*(3), 427-444. https://doi.org/10.1007/s11858-018-0916-6

Brousseau, G. (2007). *Iniciación al estudio de la teoría de las situaciones didácticas*. Libros del Zorzal.

Buck, A. J., Gross, M., Hakim, S., & Weinblatt, J. (1993). Using the Delphi process to analyze social policy implementation: A post hoc case from vocational rehabilitation. *Policy Sciences*, *26*, 271-288. https://doi.org/10.1007/BF00999473

Cabero, J. (2014). Formación del profesorado universitario en TIC. Aplicación del método Delphi para la selección de los contenidos formativos. *Educación XX1, 17*(1), 111-132. https://doi.org/10.5944/educxx1.17.1.10707

Cabero, J., & Barroso, J. (2013). La utilización del juicio de experto para la evaluación de TIC: el coeficiente de competencia experta. *Bordón. Revista de Pedagogía, 65*(2), 25-38. https://doi.org/10.13042/brp.2013.65202

Cabero, J., & Infante, A. (2014). Empleo del método Delphi y su empleo en la investigación en comunicación y educación. *EDUTEC. Revista electrónica de tecnología educativa*, *48*, a272-a272. https://doi.org/10.21556/edutec.2014.48.187

Cabero, J., & Llorente, M. C. (2013). La aplicación del juicio de experto como técnica de evaluación de las tecnologías de la información y comunicación (TIC). *Revista de Tecnología de Información y Comunicación en Educación*, *7*(2), 11-22.

Charalambous, C.Y., & Praetorius, A.K. (2018). Studying mathematics instruction through different lenses: setting the ground for understanding instructional quality more comprehensively. *ZDM Mathematics Education*, *50*(3), 355–366. https://doi.org/10.1007/s11858-018-0914-8

Clayton, M. J. (1997). Delphi: a technique to harness expert opinion for critical decision-making tasks in education. *Educational psychology*, *17*(4), 373-386. https://doi.org/10.1080/0144341970170401

Diego-Mantecón, J. M., & Córdoba-Gómez, F. J. (2019). Adaptación y validación del MRBQ (Mathematics-Related Beliefs Questionnaire) al contexto colombiano con estudiantes de secundaria. *Educación matemática*, *31*(1), 66-91. https://doi.org/10.24844/em3101.03

Dolores, C., & Ibáñez, G. (2020). Conceptualizaciones de la Pendiente en Libros de Texto de Matemáticas. *Bolema: Boletim de Educação Matemática*, *34*, 825-846. https://doi.org/10.1590/1980-4415v34n67a22

Duval, R. (1993). Registros de representación semiótica y funcionamiento cognitivo del pensamiento. En E. Hitt (Ed.), *Investigaciones en Matemática Educativa II* (pp. 173-201). Grupo Editorial Iberoamérica.

Eccius-Wellmann, C., Lara-Barragán, A. G., Martschink, B., & Freitag, S. (2017). Comparación de perfiles de ansiedad matemática entre estudiantes mexicanos y estudiantes alemanes. *Revista iberoamericana de educación superior*, *8*(23), 69-83. https://doi.org/10.22201/iisue.20072872e.2017.23.246

Garzón, D. (2017). Análisis de las decisiones del profesor de matemáticas en su gestión de aula. *Revista Educación Matemática, 29*(3), 131-160. https://doi.org/10.24844/EM2903.05

Gordon, T. J. (1994). The Delphi method. In J.C. Glenn & T.J. Gordon (Eds.). *Futures Research Methodology Version 2.0.* (CD-

Rom). American Council for the United Nations University.

Hill, H., Blunk, M.L., Charalambous, C.Y., Lewis, J.M., Phelps, G.C., Sleep, L., & Ball, D. (2008). Mathematical Knowledge for Teaching and the Mathematical Quality of Instruction: An Exploratory Study. *Routledge Cognition and Instruction*, *26*(4). https://doi.org/10.1080/07370000802177235

Kaiser, G. (2020) Mathematical Modelling and Applications in Education. In S. Lerman (Eds.), *Encyclopedia of Mathematics Education*. Springer. https://doi.org/10.1007/978-3-030-15789-0_101

Landeta, J. (1999). *El método Delphi. Una técnica de previsión para la incertidumbre*. Ariel.

Linstone, H. A., & Turoff, M. (1975). *The delphi method*. Addison-Wesley.

Liu, R. D., Wang, J., Star, J. R., Zhen, R., Jiang, R. H., & Fu, X. C. (2018). Turning potential flexibility into flexible performance: Moderating effect of self-efficacy and use of flexible cognition. *Frontiers in psychology, 9*, 646. https://doi.org/10.3389/fpsyg.2018.00646

Llorente, L. (2013). Assessing Personal Learning Environments (PLEs). An expert evaluation. *Journal of New Approaches in Educational Research)*, *2*(1), 39-44. https://doi.org/10.7821/naer.2.1.39-44

López-Gómez, E. (2018). El método Delphi en la investigación actual en educación: una revisión teórica y metodológica. *Educación XX1*, *21*(1), 17-40. https://doi.org/10.5944/educXX1.20169

Ludwig, B. (1997). Predicting the future: Have you considered using the Delphi methodology. *Journal of extension*, *35*(5), 1-4.

Martínez, A. (1988). *Diseños experimentales: métodos y elementos de teoría*. Trillas.

Martino, J. (1999). Thirty years of change and stability. *Technological Forecasting and Social Change*, *62*(1-2), 13-18. https://doi.org/10.1016/S0040-1625(99)00011-6

Martins, G. A., & Theóphilo, C. R. (2007*). Metodologia da Investigação Científica para Ciências Sociais Aplicadas*. Atlas

Matsumura, L. C., Garnier, H., Slater, S. C., & Boston, M. (2008). Toward measuring instructional interactions 'at-scale'. *Educational Assessment*, *13*(4), 267–300. https://doi.org/10.1080/10627190802602541

Maz-Machado, A., Madrid, M. J., León-Mantero, C., & Jiménez-Fanjul, N. (2019). Mathematical practical sessions with manipulatives: Trainee teachers' perceptions of their utility. *South African Journal of Education*, *39*(2), Art. #1620. https://doi.org/10.15700/saje.v39ns2a1620

McMillan, S.S., King, M., & Tully, M.P. (2016). How to use the nominal group and Delphi techniques. *International Journal of Clinic Pharmacy*, *38*, 655–662. https://doi.org/10.1007/s11096-016-0257-x

Mengual, S., Roig, R., & Mira, J.B. (2016). Delphi study for the design and validation of a questionnaire about digital competences in higher education. *International Journal of Educational Technology in High Education*, *13*, 12. https://doi.org/10.1186/s41239-016-0009-y

Molina, O., & Samper, C. (2019). Tipos de problemas que provocan la generación de argumentos inductivos, abductivos y deductivos. *Bolema: Boletim de Educação Matemática*, *33*, 109-134. https://doi.org/10.1590/1980-4415v33n63a06

Moore, C. M. (1987). *Group techniques for idea building*. Sage Publications, Inc.

Novakowski, N., & Wellar, B. (2008). Using the Delphi technique in normative planning research: methodological design considerations. *Environment and Planning A*, *40*(6), 1485-1500. https://doi.org/10.1068/a39267

Okoli, C., & Pawlowski, S. D. (2004). The Delphi method as a research tool: an example, design considerations and applications. *Information & management*, *42*(1), 15-29. https://doi.org/10.1016/j.im.2003.11.002

Openshaw, R., & Walshaw, M. (2019). *Transnational Synergies in School Mathematics and Science Debates*. Springer Nature. https://doi.org/10.1007/978-3-030-28269-1

Orton, A. (2003). *Didáctica de las matemáticas*. Ministerio de Educación, Cultura y Deporte-Ediciones Morata.

Parenté, F. J., & Anderson, J.K. (1987). Delphi inquiry systems. En G. Wright & P. Ayton (Eds.), *Judgemental Forecasting* (pp. 129-156). John Wiley.

Patton, M. (1990). *Qualitative evaluation and research methods*. Newbury Park, CA.

Piburn, M., & Sawada, D. (2000). *Reformed teaching observation protocol (RTOP): Reference manual. Technical Report*. Arizona Collaborative for Excellence in the Preparation of Teachers.

Planas, N., Arnal-Bailera, A., & García-Honrado, I. (2018). El discurso matemático del profesor: ¿Cómo se produce en clase y cómo se puede investigar? *Enseñanza de las Ciencias*, *36*(1), 45-60. https://doi.org/10.5565/rev/ensciencias.2240

Praetorius, A. K., & Charalambous, C. Y. (2018). Classroom observation frameworks for studying instructional quality: looking back and looking forward. *ZDM Mathematics Education*, *50*(3), 535-553. https://doi.org/10.1007/s11858-018-0946-0

Rodríguez, J.M., & Ruiz, J. (2019). El clima social en centros educativos: percepción del profesorado de Educación Secundaria Obligatoria de la Comunidad de Madrid. *Revista de Investigación Educativa*, *37*(1), 231-250. https://doi.org/10.6018/rie.37.1.320541

Rowland, T., Turner, F., Thwaites, A., & Huckstep, P. (2009). Transformation: Using examples in mathematics teaching. In *Developing Primary Mathematics Teaching: Reflecting on Practice with the Knowledge Quartet* (pp. 67-100). SAGE Publications. https://doi.org/10.4135/9781446279571

Schlesinger, L., & Jentsch, A. (2016). Theoretical and methodological challenges in measuring instructional quality in mathematics education using classroom observations. *ZDM Mathematics Education*, *48*, 29–40. https://doi.org/10.1007/s11858-016-0765-0

Schneider, M., Rittle-Johnson, B., & Star, J. R. (2011). Relations between conceptual knowledge, procedural knowledge, and procedural flexibility in two samples differing in prior knowledge. *Developmental Psychology*, *47*(6), 1525-1538. https://doi.org/10.1037/a0024997

Shah, H., & Kalaian, S. A. (2009). *Which Parametric Statistical Method to Use For Analyzing Delphi Data? Journal of Modern Applied Statistical Method*, *8*(1), 226-232. https://doi.org/10.22237/jmasm/1241137140

Teddlie, C., Creemers, B., Kyriakides, L., Muijs, D., & Yu, F. (2006). The international system for teacher observation and feedback: Evolution of an international study of teacher effectiveness constructs. *Educational Research and Evaluation*, *12*(6), 561–582. https://doi.org/10.1080/13803610600874067

Thomas, C. A., & Berry III, R. Q. (2019). A Qualitative Metasynthesis of Culturally Relevant Pedagogy & Culturally Responsive Teaching: Unpacking Mathematics Teaching Practices. *Journal of Mathematics Education at Teachers College*, *10*(1), 21-30. https://doi.org/10.7916/jmetc.v10i1.1668

Thompson, C. J., & Davis, S. B. (2014). Classroom observation data and instruction in primary mathematics education: Improving design and rigour. *Mathematics Education Research Journal*, *26*(2), 301–

323. https://doi.org/10.1007/s13394-013-0099-y

van den Ham, A. K., & Heinze, A. (2018). Does the textbook matter? Longitudinal effects of textbook choice on primary school students' achievement in mathematics. *Studies in Educational Evaluation*, *59*, 133-140. https://doi.org/10.1016/j.stueduc.2018.07.005

Van Zoest, L. R., Peterson, B. E., Rougée, A. O., Stockero, S. L., Leatham, K. R., & Freeburn, B. (2021). Conceptualizing important facets of teacher responses to student mathematical thinking. *International Journal of Mathematical Education in Science and Technology*, 1-26. https://doi.org/10.1080/0020739X.2021.1895341

Vergnaud, G. (2013). Pourquoi la théorie des champs conceptuels? *Infancia y Aprendizaje*, *36*(2), 131-161. https://doi.org/10.1174/021037013806196283

von der Gracht, H.A. (2012). Consensus measurement in Delphi studies: review and implications for future quality assurance. *Technological Forecasting and Social Change*, *79*(8), 1525-1536. https://doi.org/10.1016/j.techfore.2012.04.013

Walkowiak, T. A., Berry, R. Q., Meyer, J. P., Rimm-Kaufman, S. E., & Ottmar, E. R. (2014). Introducing an observational measure of standards-based mathematics teaching practices: Evidence of validity and score reliability. *Educational Studies in Mathematics*, *85*(1), 109–128. https://doi.org/10.1007/s10649-013-9499-x

Weber, K. E., Gold, B., Prilop, C. N., & Kleinknecht, M. (2018). Promoting pre-service teachers' professional vision of classroom management during practical school training: Effects of a structured online-and video-based self-reflection and feedback intervention. *Teaching and Teacher Education*, *76*, 39-49. https://doi.org/10.1016/j.tate.2018.08.008

Authors / Autores

Arteaga-Martínez, Blanca (blanca.arteaga@edu.uned.es) 0000-0002-1079-1526

Blanca Arteaga has a Degree in Mathematics (Universidad Autónoma de Madrid) and PhD in Education (Universidad Complutense de Madrid). Dr. Arteaga has combined her work as primary and secondary teacher with university lecturing (Universidad Carlos III de Madrid, Saint Louis University in Madrid, Universidad Internacional de La Rioja, Universidad de Alcalá and Universidad Rey Juan Carlos), research and project management. She is currently Assistant Professor at the Universidad Nacional de Educación a Distancia (UNED). Her research has focused on action-research in the classroom, with the didactics of mathematics as the backbone. In that areaa, she has managed and participated in several funded research projects and educational innovation projects with different universities and school networks. Dr. Arteaga is a member of the Adaptive Pedagogy research group at the Complutense University of Madrid.

Macías-Sánchez, Jesús (j.macias@ucm.es) 0000-0001-9798-7654

Degree in Mathematics from the Universidad Complutense of Madrid, Master in Teacher Training and Master in Advanced Studies in Pedagogy from the Universidad Complutense of Madrid and PhD in Education from the Universidad Complutense of Madrid. He combines his work as a teacher at compulsory levels with teaching at university (Universidad Internacional de La Rioja, Universidad Europea de Madrid and Universidad Complutense of Madrid). Currently he is adjunct professor at the Universidad Complutense of Madrid (UCM) and teacher in compulsory education at the Colegio Brotmadrid. His research has focused on action research in the classroom, based on the teaching-learning process of mathematics. He has participated in some funded research projects and educational innovation projects.

Pla-Castells, Marta (marta.pla@uv.es) 0000-0002-2088-2536

Degree in Mathematics from Universitat de València (UVEG) and PhD in Computer Science from the same university. She has worked as director of European projects for the Institute of Robotics and ICT of the UVEG while she was associate professor in departments of the Faculty of Mathematics of the UVEG and the Universitat Jaume I of Castelló (UJI). At present she is a lecturer at UVEG. Due to her experience in software programming, she has been able to apply new technologies in the field of mathematics education. She has also worked in continuous training of primary school teachers in collaboration with the centre for training, innovation, and resources for teachers of the Valencian Regional Government. She currently belongs to the Mathematical Modelling research group at the Department of Didactics of Mathematics at the UVEG.

Ramírez-García, Mónica (mramirez@lasallecampus.es) 0000-0002-1198-2017

Degree in Mathematical Sciences, from the Universidad Autónoma de Madrid (Spain) in 1998 and PhD in Education from the Universidad Complutense of Madrid (Spain). She has worked as a professor at the Department of Experimental, Social and Mathematical Didactics in the Faculty of Education at the Universidad Complutense de Madrid (Spain) since 2011 and she currently works at the Centro Superior de Estudios Universitarios La Salle (Spain) since 2006. Her research interests have focused on the teaching-learning of mathematics in the early childhood education. In 2016, she started collaborating with several researchers from the Universidad de Sevilla and the Universidad de Huelva, applying the Mathematics Teachers Specialized Knowledge model (MTSK) both to her activity as a mathematics teacher educator and as a researcher.

RELIEVE [ISSN: 1134-4032]

Revista **EL**ectrónica de **I**nvestigación y **EV**aluación **E**ducativa

E-Journal of Educational Research, Assessment and Evaluation