Journal on Mathematics Education
Volume 14, No. 4, 2023, pp. 781-796

Development of a protocol to measure mathematics higher-order thinking skills in Mexican primary schools

Lesly Yahaira Rodriguez-Martinez[1], Paul Hernandez-Martinez[2,*], Maria Guadalupe Perez-Martinez[3]

[1]Centro de Investigación y Docencia Económicas, Aguascalientes, Mexico
[2]Department of Mathematics, Swinburne University of Technology, Melbourne, Australia
[3]Department of Education, Universidad Autónoma de Aguascalientes, Aguascalientes, Mexico
*Correspondence: phernandezmartinez@swin.edu.au

Received: 20 June 2023 | Revised: 5 October 2023 | Accepted: 10 October 2023 | Published Online: 19 October 2023

Abstract

This paper aims to describe the development process of the Observation Protocol for Teaching Activities in Mathematics (POAEM) and to report the findings from the qualitative and statistical analyses used to provide evidence of validity and reliability of the information collected with the first version of the POAEM. As part of this development process, 20 teachers from Mexican primary schools were videotaped twice while teaching mathematics. The study assessed the reliability of the POAEM rubrics. Results showed that the dimensional structure of the instrument can be grouped in one factor. A generalizability study provided information on the different sources of error in the measurement, showing that the dimensions accounted for 78% of the variance. This study provides an exemplar of the design and validation of an instrument that can help other researchers develop their own instruments and data collection to generate evidence of validity and reliability in different sociocultural contexts.

Keywords: Educational Assessment, Higher Order Thinking Skills, Observation Protocol, Primary School Mathematics Education, Teaching Mathematics

How to Cite: Rodriguez-Martinez, L. Y., Hernandez-Martinez, P., & Perez-Martinez, M. G. (2023). Development of a protocol to measure mathematics higher-order thinking skills in Mexican primary schools. *Journal on Mathematics Education, 14*(4), 781-796. http://doi.org/10.22342/jme.v14i4.pp781-796

Studies in Mathematics Education around the world have adopted different research approaches to analyze teaching activities to create learning opportunities for students. For example, researchers have designed interventions (Stylianides & Stylianides, 2013), new teaching models (Burkhardt, 2006; Cevikbas & Kaiser, 2020), and instruments to measure students' learning gains within those practices (Boston, 2012; Schlesinger & Jentsch, 2016; Turner et al., 2018).

In Mexico, the evaluation of teaching activities has been carried out mainly through characterizing teachers' practices, beliefs, and conceptions, analyzing teachers' mastery of the content, and student attainment on large-scale tests (Ávila et al., 2013; López & Mota, 2003; Martínez Rizo & Chávez, 2015). However, there is a lack of studies in the Mexican context in particular – and the Hispano-American context in general –focusing on developing research instruments that allow teachers and educational researchers to obtain quality information about the efficacy of teaching activities on students' learning. Furthermore, of the existing studies, not many report evidence of the validity and reliability of the data

collected (Ávila et al., 2013).

Hence, this paper aims to contribute to filling this gap by describing a methodological study that designed and validated an instrument for the analysis of mathematics teaching activities in the context of primary school lessons in Mexico. Specifically, we report on the process of developing a lesson observation protocol and the validation of the information obtained from it. We called this instrument the Observation Protocol for Teaching Activities in Mathematics (POAEM in Spanish: Protocolo de Observación para Evaluar las Actividades de Enseñanza en Matemáticas).

With this paper, we seek to promote further development and use of such instruments in the Hispano-American context and provide a model for those who wish to develop their own instruments, suitable to their sociocultural contexts. An important feature of our work is the combined use of sociocultural and psychometric approaches, since we argue that the context in which an instrument is designed and validated is of utmost importance. Therefore, while our contribution is primarily aimed at Mexican/Hispano-American researchers and practitioners, this paper will also benefit others worldwide interested in developing similar instruments.

This paper will answer the following research questions:

- To what extent does the information collected with the observation protocol appear to be a valid measure?
- What is the reliability of the observation measures of the teaching activities based on the designed observation protocol?
- What is the factorial structure of the observation protocol to evaluate teaching activities in mathematics?
- What is the variability of the data obtained with the designed observation protocol? For example, raters, observation occasions, teachers, and school context, among other facets of error.

Research on Teaching Activities

Learning in the classroom greatly depends on the kind of intellectual work generated through teaching activities and other learning opportunities in which students are involved (NCTM, 2014; Picaroni & Loureiro, 2010; Zolkower & Bressan, 2012). A teaching activity is a system integrated by different elements and characteristics derived from the interactions between teachers and students and among students.

A large body of literature has investigated features of teaching activities and practices that promote student learning, acknowledging teaching has influence on student learning. For example, some scholars have emphasized the importance of establishing clear learning goals, giving students opportunities to work with different levels of cognitive demand, contextualizing teaching activities, and promoting students' self-assessment as practices that are most likely to influence students' learning (Aubuson et al., 2014; Llinares, 2008; NCTM, 2014; Picaroni & Loureiro, 2010; Swan, 2015; Zolkower & Bressan, 2012). So, the intellectual work proposed by teachers through teaching activities and the opportunities that these represent for the construction of new knowledge can stimulate reasoning and involve students in developing higher order thinking skills.

The literature review on which our instrument was based (Rodríguez-Martínez, 2018) aimed to identify features of teaching activities and practices that promote higher order thinking skills. Higher order thinking skills include "analysis, comparison, inference, interpretation, evaluation, and synthesis applied in academic fields and problem-solving contexts" (Osman, 2013, p. 1598). Our interest in higher order

thinking skills in mathematics education stemmed from research on mathematical tasks and activities, which has found that using higher-order thinking skills in the classroom increases students' learning. Still, not all classroom activities provide opportunities for students to think this way (NCTM, 2014). In addition, research has shown that these types of activities are "the most difficult to implement" (NCTM, 2014, p. 17), and even when teachers design activities to promote higher-order thinking skills, many times when implemented in the classroom, these are transformed into less demanding tasks (Boston & Smith, 2009).

A teaching activity entails actions aimed at promoting specific types of learning. Therefore, teaching activities indicate the learning opportunities provided to students in classroom settings since students build meanings and knowledge according to their own experiences (Boston & Wolf, 2006; Donovan & Bransford, 2005).

Many authors agree that teaching practices should be supported by activities that encourage intellectual challenges and that these activities should be relevant and valuable for students to achieve deep learning (Aubuson et al., 2014; Gulikers et al., 2004; McTighe & Wiggins, 2012; Newmann et al., 1996; Newmann & Wehlage, 1993). These characteristics indicate the importance of teaching activities as an input for learning. Especially those activities that are attractive to students because of their personal and social value, activities that mean intellectual challenges, and activities that require the application of previous knowledge (Aubuson et al., 2014; McTighe & Wiggins, 2012; Newmann et al., 1998).

Learning Achievements in Mathematical Thinking at Primary Education in Mexico

Results of national and international achievement tests related to mathematics show that most Mexican primary school students do not perform at the highest levels of proficiency, which implies that they need to develop mathematics higher order thinking skills. For example, the Fourth Regional Comparative and Explanatory Study (ERCE) (UNESCO, 2021), applied to third and sixth-grade students in 16 Latin American countries in 2019, found that only 8.9% of third-year students and 11.9% of sixth-year students achieved the highest level of proficiency in mathematics, a level where it is expected students to be developing higher order thinking skills. The National Plan for the Assessment of Learning (PLANEA) (INEE, 2018), a Mexican test that measured students' key learning achievements in, amongst other fields, mathematics, found that only 8% of them performed at the highest level. These results mean that most students who complete primary education have not developed high cognitive skills that allow them to explore and understand the nature of mathematical concepts to self-regulate their cognitive processes.

The Program for International Student Assessment (PISA) (OECD, 2015, 2019), assessed the skills and knowledge of 15-year-old students in Science, Reading and Mathematics and how students can apply what they learned in school to real-life situations. Unfortunately, PISA 2018 results in mathematics showed that only 44% of Mexican students scored at level 2 or higher, while the average in OECD countries was 76%. Furthermore, only less than 1% of Mexican students scored at the highest levels of proficiency (level 5 and higher), while the average of OECD countries was 11%. So, less than 1% of Mexican students take an active and creative position in their approach to mathematical problems. This is worrisome because the Mexican Educational System seems unable to provide learning opportunities for many students to develop lifelong competences.

METHODS

Design of the POAEM

Literature in teaching practices suggests different approaches to gathering information about them. For example, information given by the subjects (questionnaires and scales, diaries, interviews, etc.); and instruments to analyze academic products such as portfolios and observation protocols (Martínez Rizo, 2012).

Observation protocols are the primary approach to obtaining information on teaching practices. The Measure of Effective Teaching Project (MET) used different observation protocols to measure teaching effectiveness. The MET project was based on a study that sought to analyze measures of teacher effectiveness to establish the characteristics of effective teaching practices and provide helpful knowledge about teacher performance and professional development (Met Project, 2010). However, our epistemological position is that the sociocultural context of these protocols makes them unsuitable for use in the Mexican context.

A sociocultural pedagogic approach confers importance on social interactions among teachers, students, and contextual elements of the classroom (e.g., resources available, socio-normative rules). This approach is essential in developing different educational measurement instruments since interactions and context influence decisions about what to measure, the purpose, and the interpretation of results. It is particularly important in developing content specifications, ensuring that the test content is related to the construct intended to measure and that the construct is relevant. Thus, we integrate sociocultural and psychometric approaches as the basis for developing the POAEM rubrics since both approaches allow us to collect sources of validity evidence and support inferences derived from applying the rubrics. Psychometric methods are needed to assess reliably and to address the contexts in which children learn. For us, learning contexts comprise the characteristics of the Mexican context, specifically the national curriculum for primary education, free textbooks, and the initiatives promoted to improve the teaching and learning of mathematics in this country due to the importance of addressing cultural and contextual differences in Mexico.

We establish features of teaching activities and practices that promote student learning, acknowledging the influence that teaching has on student learning. Given these features, the POAEM's main purpose is to provide information about what teachers and students do in a mathematics lesson to promote higher-order thinking skills. In doing so, we can increase the knowledge to better understand the variation in learning, enriched teacher preparation, and design educational improvement initiatives, since in Mexico, there is a need to document opportunity gaps where significant differences in SES and ethnicity predominate (Jensen et al., 2016).

In order to design useful measures, we based our new protocol on dimensions consistent with international protocols but adjusting them to the purposes and standards of the Mexican national curriculum. The purpose was to design an observation protocol that would allow collecting valid and reliable information on teaching activities in the Mexican context. Consideration is given to the great inequality of socioeconomic, geographical and technological conditions and the low results in mathematics reported by standardized tests.

Some of the observation protocols that were reviewed as a background for the design of the POAEM are: a) the Mathematical Quality of Instruction (MQI), designed to measure the mathematical content available to students during instruction; b) the Mathematics Classroom Observation Protocol for Practices (MCOP), designed to measure the practices within the mathematics classroom for teaching

lessons that are goal-oriented toward conceptual understanding; c) the Mathematics Scan (M-Scan), designed to measure the extent to which the dimensions: structure of the lesson, multiple representations, students' use of Mathematical tools, cognitive demand, mathematical discourse community, explanation and justification, problem solving and connections and applications, are present in the math lessons; and d) the Mathematics Integrated into Science: Classroom Observation Protocol (MISCOP), a protocol aimed at defining the level of integration of mathematics and science in the classroom (Hill et al., 2012; Gleason et al., 2017; Walkowiak et al., 2013; Judson, 2013).

The design process and implementation of the different observation protocols developed from the MET project are a reference for new methodological studies such as the one presented in this paper. However, not many studies of classroom observation protocols reported in the literature make explicit the validation process of the information collected with those instruments. Bostic et al. (2021) found that from 114 published studies only 39% of those mentioned a formalized classroom observation protocol. Furthermore, they found that validity evidence for classroom observation protocols "was rarely found within the manuscripts themselves" (Bostic et al., 2021, p. 17). Hill and Shih (2009) previously reported that only 17% of quantitative studies published in the Journal for Research in Mathematics Education (the top journal in the field) employed measures with validity evidence or reported their own validity evidence. Hence, in this paper we explicitly present the phases to build the POAEM and the process of validating the information collected with it: a) theoretical framework and construct operationalization, b) feedback from a group of expert judges, and c) testing of the POAEM.

Aspects of the Reliability of the Observation Measures of the Teaching Activities based on the Designed Observation

A seven-dimensions framework was developed from a literature review (Rodríguez-Martínez, 2018) to create the POAEM. We reviewed both, literature focused on mathematics teaching and learning and literature without a specific subject orientation. Attention was focused on the coincidences in different studies on teaching activities since they are considered to represent basic elements from which it is possible to establish a reference for designing higher order thinking skills tasks. Each dimension proposed became a rubric which was expected to be a tool that provides information so that teachers can improve their own practices. All rubrics use a 4-point scale (1 = low level through to 4 = high level), this rating scheme facilitates comparisons across dimensions and enables the classroom observer to develop a strong idea of what each score level refers to in a classroom situation (the protocol and the full description of the dimensions and rating scheme can be accessed at https://figshare.com/s/86f82711f3b4bfc54108). Table 1 summarizes the seven rubrics and their main components, with the main references used in their design.

The POAEM theoretical framework was derived initially from the literature review and each dimension was discussed in two working sessions with instrument design methodologists and a committee of experts who analyzed the relevance of each dimension and rubric for the Mexican context. The expert committee consisted of a primary school teacher, two educational researchers specialized in teaching mathematics, and two educational researchers specialized in instrument design, all with extensive experience and expertise in the Mexican primary school system.

The first version of the POAEM was reviewed in two four-hour sessions. The expert judges discussed in pairs their observations and later in plenary. Experts' opinions allowed us to ensure that the identified dimensions represented the variables of interest. For qualitative analysis, each expert assessed the seven dimensions from the theoretical framework, keeping in mind three key elements: a) usefulness

of the information and suitability for Mexican primary school classrooms, b) clarity of ideas in each dimension, and c) information that could be removed from the theoretical framework. Feedback from each judge was discussed in collaborative work with the other judges.

Table 1. POAEM theoretical framework

Dimensions	Description	Elements	Main references
Clarity of the task	What students should know and be able to do after specific tasks.	Learning goals Purpose of the task	NCTM, 2014; Newmann et al.,1998
Task cognitive demand	Kind and level of thinking required to solve tasks.	Memorization, procedures without connections, procedures with connections, doing mathematics.	Stein et al., 1996
Resources used	Use of resources to promote the learning process.	Making sense about mathematical ideas Think in a mathematical way Communicate mathematical thinking.	NCTM, 2014
Alternative solutions	Flexibility to promote and accept a variety of strategies to solve mathematical problems.	Multiple ways of solving a problem Critical awareness and a mathematical reasoning.	Picaroni & Loureiro, 2010; Wiggins, 1998
Questions for reflection	Type of questions posed.	Self-reflection Justify and demonstrate answers	Donovan & Bransford, 2005; Hiebert et al., 1997; NCTM, 2014
Connection / application of mathematical knowledge	The context as a problematic situation.	Application of mathematical knowledge	Aubuson et al., 2014; McTighe & Wiggins, 2012
Collaborative work	Interaction, negotiation and cooperation among students.	Development of common ideas, procedures or products Share ideas or procedures to build or strengthen some mathematical knowledge.	NCTM, 2014; Secretaría de Educación Pública, 2011

For example, the judges considered the Mexican national strategy in Primary Mathematics Education, which consists mainly of mathematical challenges. The curriculum emphasizes the importance of implementing tasks that promote meaningful and in-depth learning for students (Secretaría de Educación Pública, 2011). The government has produced an official textbook series called “Desafíos Matemáticos” (Mathematical Challenges), with the aim of supporting the accomplishment of curriculum objectives by providing “both teacher and students with attractive, useful, mysterious and clever mathematical challenges” to solve (Secretaría de Educación Pública, 2012, p. 7). Therefore, the judges made sure that the dimensions of the POAEM aligned with what could be observable in the context of Mexican primary classrooms according to the national curriculum specifications.

A further example of how the sociocultural context shaped the dimensions of the POAEM was when the judges discussed the dimension “Clarity of the task”. Part of this discussion focused on the pertinence of communicating to students the learning objectives before allowing them to work on the task,

because this might act against the challenge that should be intrinsic to the task, where students need to struggle to actively construct meaning, or against dealing with the complexity that is inherent to activities that aim to develop higher order thinking skills.

The results of this feedback helped us to improve the first POAEM version: we adjusted the criteria of each dimension, a new level of performance was added, we included a brief description for every dimension, and we improved the format to assess each dimension. A second POAEM version was reviewed with the same strategy and by the same team of expert judges, who suggested to adjust the description of the performance levels in the rubrics in order to test the protocol on videorecorded lessons segmented by tasks.

We then recruited 20 fifth-grade primary school teachers who worked in public schools. Due to the characteristics of the study, a representative sample was not required, but only the number and type of subjects necessary to carry out the appropriate statistical tests. Two mathematics lessons, taught on different days, were video recorded for each teacher. Therefore, forty lessons were video recorded, each lasting an average of 50 minutes, and all recordings were segmented into mathematical tasks. A mathematical task could "range from a set of routine exercises to a complex and challenging problem that focuses students' attention on a particular idea" (NCTM, 2014, p. 19). Segmentation involved the analysis of each video to identify the tasks it comprised. The number of tasks per lesson varied from two to four.

RESULTS AND DISCUSSION

Dimensions and criteria on the POAEM were tested through statistical analyses. In order to perform the statistical analyses, we used a database with scores awarded by two trained observers; both observers used the performance levels from the model (rubrics) to assess the 40 videos of mathematics classes. The 40 videos were segmented in tasks, according to the definition presented above; in total the videos comprised 111 mathematical tasks. In the following sections we will present the results of the statistical analysis of the scores awarded to the mathematical tasks.

Exploratory and Confirmatory Factor Analysis

Exploratory Factorial Analysis (EFA) and a Confirmatory Factorial Analysis (CFA) were performed to reach the number and composition of factors that explain the common variance of the set of dimensions that conform the theoretical framework; as well as to locate isolated dimensions, that is, dimensions that do not belong to the categories assessed by this theoretical framework (Lloret et al., 2014). Factor analysis assumes that "the observed variables are indicators of a number of common latent factors or variables ... and the item (or in this case the dimension) is a manifestation of this factor" (Lloret et al., 2014, p. 1152). Therefore, different authors agree to use the factor analysis to report the dimensionality of a structural model as well as Cronbach alpha for the reliability of the model (Basto & Pereira, 2012; Courtney & Gordon, 2013). For this theoretical model, the first step was to verify the rank of adequacy to carry out the factorial analysis from the KMO test (Kaiser, Meyer and Olkin) and Bartlett's sphericity. Both analyzes showed adequate values to perform the EFA, since the KMO adequacy measure was greater than .50 and the value of significance in the sphericity test was less than .05 (Table 2).

Table 2. KMO test and Bartlett sphericity test

KMO		.816
Bartlett sphericity test	Chi-square	204.295
	gl	21
	Sig.	.000

Once the feasibility to perform the factor analysis was verified, we used the "R Factor program" for the extraction, rotation, and selection of factors into the factor analyzes through the SPSS package. The extraction method was the factorization of main axes, with an oblique rotation (Quartimin), since Basto and Pereira (2012) have stated that this type of rotation can generate more reliable solutions. Table 3 presents the factorial loads for the seven dimensions confirmed by the theoretical framework. The results showed that the load excludes the "clarity of the task" dimension from the first factor. This means that "clarity of the task" dimension could be part of another different category from the one that groups the remaining six dimensions.

Table 3. Exploratory Factorial Analysis

Factor	**(111 cases)**	
	F1	**F2**
Clarity of the task	.449	**.682**
Cognitive demand	**.746**	-.034
Resources used	**.658**	.282
Alternative solutions	**.668**	-.508
Questions for reflections	**.769**	-.018
Connection of mathematical knowledge	**.554**	-.497
Collaborative work	**.788**	.205

To determine the correct number of factors to be retained according to the EFA, we used different extraction methods: Kaiser rule (eigenvalues), parallel analysis, optimal coordination, acceleration factor, and Velicer test. The results indicate that Kaiser's rule (eigenvalues) retains two factors, while parallel analysis sets the cut-off point by one factor, as well as the optimal coordination, the Velicer test, and the acceleration factor (Figure 1).

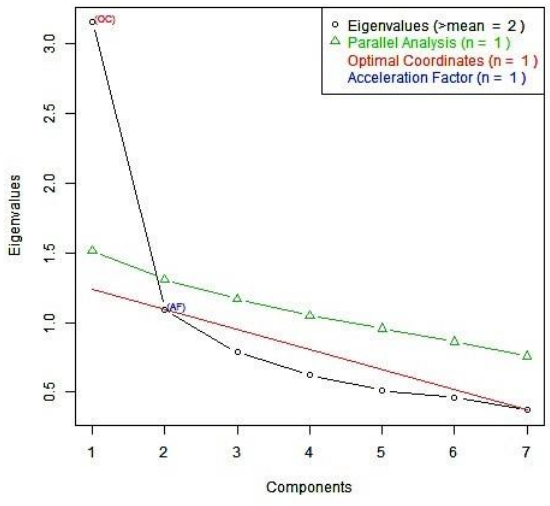

Figure 1. Estimation of factors through four methods. Source: own design based on database

The variability of the data regarding each factor shows that the theoretical model includes the seven proposed dimensions, and it explains 60% of the variance with two factors; however, if the dimension “Clarity of the task” is eliminated, the model can explain 50% of the variance with a single factor. This is summarized in Table 4.

Table 4. Explained variance and eigenvalues

AFE (seven dimensions)				AFE (six dimensions)			
Factor	**eigenvalues**	**% variance**	**% accumulated**	**Factor**	**eigenvalues**	**% variance**	**% accumulated**
1	3.155224	45.074629	45.074629	1	3.009643	50.160711	**50.160711**
2	1.091879	15.598273	**60.672901**	2	0.923707	15.395109	65.555820
		GFI: .899				**GFI: .899**	

The resulting model of the EFA was contrasted with a CFA by using the statistical program “R” and the libraries "Psych" and "Sem." The results showed that the dimensional structure of the model groups most of the dimensions into one factor. The assumptions to accept in a unifactorial model are met since the GFI = 0.96, and CFI = 0.96 both present an excellent adjustment index, and the approximate error rate RMSEA = 0.06 is slightly lower than the reference value (0.08). This confirms the identified structure with the AFE (Figure 2). In addition, the internal consistency of the model obtained a Cronbach's alpha of 0.78 (acceptable).

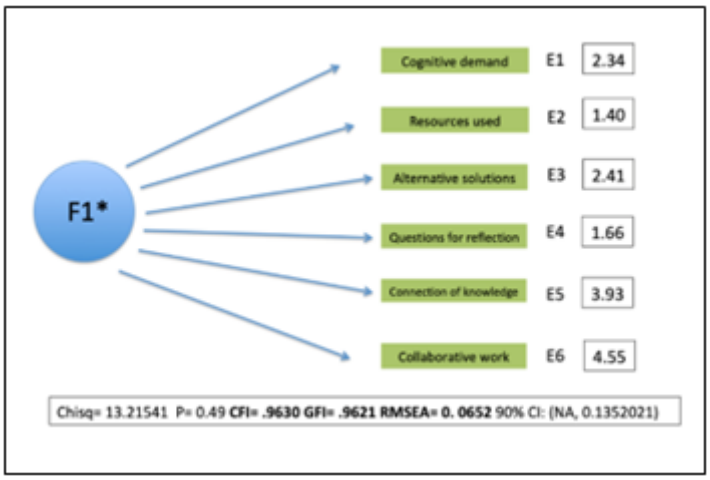

Figure 2. Unifactorial model. Source: own design based on the modifications of parameters

Both results of EFA and CFA suggest that most dimensions of the proposed model are associated with the development of higher-order thinking skills, except the "Clarity of the task." This result was expected since the purpose of “Clarity of the task” dimension is to identify how the teacher explores whether the students have understood what is requested of them, and this is measured by observing the way in which the teacher seeks to clarify the work students should do, without necessarily presenting the objective or the expected learning purpose. So, in this theoretical model, the dimension "Clarity of the task" has been included as a previous step for the promotion of the development of higher-order thinking skills since NCTM (2014), Oura (2001), Stone et al. (2008) and Newmann et al. (1998) argue that

students' understanding of what is being requested as classwork or product is a fundamental step to achieve an effective teaching practice. Therefore, teachers must ensure that students clearly understand what they should do during a task.

Internal Consistency: Cronbach's Alpha

To calculate the correlation among items in the POAEM, Cronbach's alpha coefficient of homogeneity or internal consistency was obtained. The results suggest that internal consistency is acceptable with a value of 0.785 according to the criteria of Oviedo and Campo-Arias (2005) for the Cronbach's alpha coefficient: values greater than 0.90 excellent; 0.80 to 0.90 good; 0.70 to 0.79 acceptable; 0.60 to 0.69 questionable; 0.59 to 0.50 poor; and less than 0.50 unacceptable. This means that the items that make up the instrument provide information about the same construct (teaching activities). Therefore, we have statistical evidence that the POAEM collects information consistent with what it seeks to collect.

Generalizability Study

Generalizability studies examine multiple sources of error in a measurement in order to determine the consistency of the results and the conditions under which sound levels of reliability can be obtained (Shavelson & Webb, 1991). In this study, the true score is P, which refers to the observed teacher, and the sources of variation are: the rater "R", is the degree to which a single rater's scores are systematically different from the average of all raters; the observation occasion (O), refers to the difference between the scores assigned in the two different occasions in which the same teacher was recorded; and the task segment within each occasion (S: O), which denotes the degree to which individual segments deviate from the day mean within the same teacher. This design is represented as: P x R x O x S: O. According to the tasks in which each video was segmented, this is classified as an unbalanced design "since the number of levels is unequal, this is particularly due to the differences in segments per occasion (S: O)" (Shavelson & Webb, 1991, p. 610), which varied from two to four, as mentioned above. We also included in the model the interaction effects between sources of error: teacher per day (PO); teacher per segment (PS:O); teacher per rater (PR); day per rater (OR); rater per segment (Day) (RS:O); teacher per day per rater (POR); and teacher per rater per segment (Day).

Table 5 presents variance components and corresponding contributions to overall score variance with all POAEM dimensions (column A), excluding the "clarity of the task" dimension (column B). The "clarity of the task" dimension was excluded because the AFE and the AFC suggest that most dimensions of the proposed model are associated with the development of higher order thinking skills, except the "clarity of the task".

Table 5 reveals that overall variance was greatest when all POAEM dimensions were included (78.249 versus 70.366). Regarding the error variance components, we found that the highest percentage of variance in both designs was concentrated on the differences between raters when rating different teachers (PR, 31% and 31.6%), followed by the variation on teacher practice between segments (PS, 21.9% and 22.7%). Overall, the first design reports a residual variance (PRS: O, e) of 22%, so seven-dimensional design explains 78% of the model variance (first column); while the second design reports a residual variance (PRS: O, e) of 23%, explaining 77% of the total variance of the model (second column). That is 1% less than the design with all dimensions. However, the magnitude of the error associated with the measurement facets remains low in both designs, therefore, as a whole, the dimensions of the POAEM provide information on what is to be measured. So, the smaller the error associated with the measurement facets (PRS: O, e), the greater the contribution made by the facets included in this model.

With this data we can identify if one judge is more severe than another, identify if the judges rate differently on each occasion, if a judge is rating differently in the segments within the occasions, etc.

Table 5. POAEM variance

	[A] All POAEM Dimensions		[B] Without Clarity of Task dimension	
	σ^2	%	σ^2	%
Teacher (P)	23.085	29.5%	19.497	27.7%
Occasion (O)	-.725	-0.9%	-.627	-0.9%
Segment (Day); (S:O)	.000	0.0%	.000	0.0%
Rater (r)	-.306	-0.4%	.107	0.2%
Teacher * Day (PO)	-1.902	-2.4%	-.677	-1.0%
Teacher * Segment (Day); (PS:O)	17.170	21.9%	15.163	21.5%
Teacher * Rater (PR)	24.258	31.0%	22.231	31.6%
Day * Rater (OR)	.763	1.0%	.529	0.8%
Rater * Segment (Day); (RS:O)	-1.539	-2.0%	-1.432	-2.0%
Teacher * Day * Rater (POR)	.347	0.4%	-.427	-0.6%
Teacher * Rater * Segment (Day), Error (PRS:O, e)	17.098	**21.9%**	16.002	**22.7%**
Total variance	78.249	100.0%	70.366	100.0%

CONCLUSION

The results presented above allowed us to answer the research questions in the following way: a) the design of the observation protocol was reviewed by expert judges who made sure that the dimensions of the POAEM were aligned with what could be observable in the sociocultural context of Mexican primary classrooms. Experts' opinions ensure that the identified dimensions represented the variables of interest (content validity evidence), b) reliability evidence was established with the Cronbach's alpha coefficient (0.785) to determine internal consistency and with a generalizability study to identify sources of measurement error, c) dimensions and criteria on the POAEM were tested through statistical analyses to identify the factorial structure. The results showed that the dimensional structure of the model groups most of the dimensions into one factor. This means that the items that make up the instrument are providing information about the same construct (teaching activities). Therefore, we have statistical evidence that the POAEM collects information consistent with what it seeks to collect (construct validity evidence).

The design process for this type of instruments is complex and lengthy, and it requires different stages of adjustment so that at some point generalizations can be sustained. Future work on the POAEM will require a review of the criteria included in the rubrics, as well as a new generalizability study with a balanced design, with the same number of tasks assessed per day. Therefore, the next step in the development of this protocol will be to make the appropriate adjustments and to obtain new evidence of the validity of the information collected with the instrument. So far, the evidence of validity of the information collected with the POAEM was established from the judgments of a committee of experts who discussed conceptual and operational elements of the instrument, as well as a Factor Analysis to establish the dimensional structure of the instrument. Reliability evidence was established with the Cronbach's alpha coefficient to determine internal consistency, and with a generalizability study to identify

sources of measurement error.

Analyzing complex constructs, such as teaching practices, requires establishing clear and systematic approaches to collect, analyze, and use the information collected. It is essential to define what and how will be analyzed, and how the results will be used (Stiggins, 2007). Here, we focused on what is going to be analyzed. Each dimension proposed in the POAEM is coherent with some international instruments to assess teaching quality. These dimensions are based on the research literature on what it means to develop higher-order thinking skills. Still, these are also shaped by the sociocultural context in which the protocol is meant to be applied. This emphasizes the fact that it would have been inappropriate for us to take an instrument developed for a different, noticeably sociocultural distinct context (e.g., the US educational system) and apply it directly to the Mexican context.

Our study has contributed to developing an instrument to assess teaching activities in the Mexican primary school context, where methodological studies could be more extensive. In the last decades, instruments developed elsewhere have been applied in Mexico with different purposes: to assess teaching quality, to account for the educational system, to identify components of effective teaching, and to promote professional development for teachers, among others. We contend that this is inappropriate and that sociocultural situated instruments should be developed. Furthermore, very few of these instruments developed elsewhere provide evidence of their validity and reliability. This is a common weakness in published methodological studies in general and in particular of those using classroom observation protocols, which is "concerning" (Bostic et al., 2021, p. 19). In contrast, this study documented the process followed for developing the observation protocol and for collecting evidence of their validity; in doing so, we want to document the systematic approach we used to collect, analyze, and use information to validate a theoretical model so that this can be useful to other researchers in developing instruments and collecting quality information to generate evidence of validity and reliability in different contexts (Ávila et al., 2013). This, in turn, can help teachers develop new practices through specific criteria and standards and motivate change to improve the quality of education.

Acknowledgments

The authors would like to extend their appreciation to the following individuals and agencies that contributed to the completion of this research: Mexican National Council of Humanities, Science, and Technology; Dr. José Felipe Martínez Fernández, who provided advice on the design of the study and the statistical analyses; Indira Viridiana Medina Mendoza y Sara Sofía Calvario Ruiz, provided support on segmenting the video recordings; the specialists who participated in judging the instrument; and the teachers who willingly participated in this study.

Declarations

Author Contribution : LYR-M: Conceptualization, Methodology, Investigation, Data curation, Writing - Original Draft, Editing, and Visualization.
PH-M: Writing - Review & Editing, and Methodology.
MGP-M: Conceptualization, Writing – Review & Editing, Resources, Validation and Supervision.

Funding Statement : This research was funded by the Mexican National Council of Humanities, Science and Technology (CONAHCYT).

Conflict of Interest : The authors declare no conflict of interest.

REFERENCES

Aubuson, P., Burke, P., Schuck, S., & Kearney, M. (2014). Teachers choosing rich tasks: The moderating impact of technology on student learning, enjoyment, and preparation. *Educational Researcher, 43*(5), 219-229. https://doi.org/10.3102/0013189X14537115

Ávila, A., Altamirano, A. C., Galindo, A. A. G., Ramos, M. T. G., López-Bonilla, G., & Ramírez, J. L. (2013). Una década de investigación educativa en conocimientos disciplinares en México. *Colección, 9* (786074), 510881.

Basto, M., & Pereira, J. M. (2012). An SPSS R-menu for ordinal factor analysis. *Journal of Statistical Software, 46*(4), 1-29. http://doi.org/10.18637/jss.v046.i04

Bostic, J., Lesseig, K., Sherman, M., & Boston, M. (2021). Classroom observation and mathematics education research. *Journal of Mathematics Teacher Education, 24*(1), 5-31.

Boston, M. (2012). Assessing instructional quality in mathematics. *The Elementary School Journal, 113*, 76-104. https://doi.org/10.1086/666387

Boston, M. D., & Smith, M. S. (2009). Transforming secondary mathematics teaching: increasing the cognitive demands of instructional tasks used in teachers' classrooms. *Journal of Research in Mathematics Education, 40*(2), 119-156.

Boston, M., & Wolf, M. (2006). *Assessing academic rigor in mathematics instruction: The development of instructional quality assessment toolkit. CSE Technical Report 672*. CRESST. http://www.cse.ucla.edu/products/reports/r672.pdf

Burkhardt, H. (2006). Modelling in Mathematics classrooms: Reflections on past developments and the future. *ZDM Mathematics Education, 38*(2), 178-195 https://doi.org/10.1007/bf02655888

Cevikbas, M., & Kaiser, G. (2020). Flipped classroom as a reform-oriented approach to teaching mathematics. *ZDM Mathematics Education, 52*(7), 1291-1305. https://doi.org/10.1007/s11858-020-01191-5

Courtney, M. G. R., & Gordon, M. (2013). Determining the number of factors to retain in EFA: Using the SPSS R-Menu v2. 0 to make more judicious estimations. *Practical Assessment, Research & Evaluation, 18*(8), 1-14. https://doi.org/10.7275/9cf5-2m72

Donovan, M., & Bransford, J. (2005). How Students Learn: History, Mathematics and Science in the Classroom. *National Research Council, Committee on How People Learn: A targeted report for teachers*. National Academies Press.

Gleason, J, Livers, S., & Zelkowski, J. (2017) Mathematics Classroom Observation Protocol for Practices (MCOP2): A validation study. *Investigations in Mathematics Learning, 9*(3), 111-129. https://doi.org/10.1080/19477503.2017.1308697

Gulikers, J. T., Bastiaens, T. J., & Kirschner, P. A. (2004). A five-dimensional framework for authentic assessment. Educational technology research and development, 52(3), 67-86. https://doi.org/10.1007/BF02504676

Hiebert, J., Carpenter, T., Fennema, E., Fuson, K., Wearne, D. Murray, H. Oliver, A., & Humen, P. (1997). *Making sense: Teaching and learning mathematics with understanding*. Hienemann.

Hill, H., Charalambous, Charalambos Y., & Kraft, M. (2012). When rater reliability is not enough: Teacher observation systems for generalizability study. *Educational Researcher, 41*(2), 56-64. https://doi.org/10.3102/0013189X12437203

Hill, H., & Shih, J. (2009). Examining the quality of statistical mathematics education research. *Journal of Research in Mathematics Education, 40*(3), 241-250.

Instituto Nacional para la Evaluación de la Educación [INEE]. (2018). *Planea resultados nacionales 2018* [Planea national results]. https://www.inee.edu.mx/images/stories/2018/planea/PLANEA06_Rueda_de_prensa_27nov2018

Jensen, B., Perez Martinez, M. G., & Aguilar Escobar, A. (2016). Framing and assessing classroom opportunity to learn: The case of Mexico. *Assessment in Education: Principles, Policy & Practice, 23*(1), 149-172.

Judson, E. (2013). Development of an instrument to assess and deliberate on the integration of mathematics into student-centered science learning. *School Science and Mathematics, 113*(2), 56–68. https://doi.org/10.1111/ssm.12004

Llinares, S. (2008). Aprendizaje del estudiante para profesor de matemáticas y el papel de los nuevos instrumentos de comunicación. *III Encuentro de Programas de Formación Inicial de Profesores de Matemáticas.* Universidad Pedagógica Nacional.

Lloret, S., Ferretes, A., Hernandez A., & Tomas, I. (2014). Exploratory Item Factor Analysis: a practical guide revised and updated. *Annals of Psychology 30*(3), 1151-1169. https://doi.org/10.6018/analesps.30.3.199361

López & Mota, Á. (2003). *Saberes científicos, humanísticos y tecnológicos: procesos de enseñanza y aprendizaje* [Scientific, humanistic and technological knowledge: teaching and learning processes]. Consejo Mexicano de Investigación Educativa.

Martínez Rizo, F. (2012). Procedimientos para el estudio de las practicas docentes: Revision de la literatura [An empirical study of the impact of formative assessment: A literature review]. *Revista Electronica de Investigacion y Evaluacion Educativa, 18*(1), 1–22.

Martínez Rizo, F., & Chávez, Y. (2015). *La enseñanza de las matemáticas y Ciencias Naturales en Educación Básica en México. Revisión de literatura* [Teaching Mathematics and Natural Sciences in Basic Education in Mexico. Literature review]. Universidad Autónoma de Aguascalientes.

McTighe, J. & Wiggins, G. (2012). *Understanding by design framework*. http://www.ascd.org/ASCD/pdf/siteASCD/publications/UbD_WhitePaper0312.pdf

MET Project (2010). *Validation Engine for Observational Protocols*. Bill y Melinda Gates Foundation.

National Council of Teachers of Mathematics [NCTM] (2014). *Principles to actions. Ensuring mathematical success for all.* NCTM.

Newmann, F., Marks, H., & Gamoran, A. (1996). Authentic pedagogy and student performance. *American Journal of Education, 104*(4), 280-312. https://doi.org/10.1086/444136

Newmann, F.M., Lopez, G., & Bryk, A.S. (1998) *The quality of intellectual work in Chicago schools: A baseline report*. Consortium on Chicago School Research.

Newmann, F., & Wehlage, G. (1993). Five standards of authentic instruction. *Educational Leadership, 50*(7), 8-12. https://www.ascd.org/el/articles/five-standards-of-authentic-instruction

Organization for Economic Co-operation and Development [OECD]. (2015). *Programa para la evaluación internacional de alumnos (PISA). Resultados 2015.* https://www.oecd.org/pisa/PISA-2015-Mexico-ESP.pdf

Organization for Economic Co-operation and Development [OECD]. (2019). *Programme for International Student Assessment (PISA). Results from PISA 2018*. https://www.oecd.org/pisa/publications/PISA2018_CN_MEX.pdf

Osman, K. (2013). Scientific Inventive Thinking Skills in Children. In E.G. Carayannis (Ed.), *Encyclopedia of Creativity, Invention, Innovation and Entrepreneurship*. Springer. https://doi.org/10.1007/978-1-4614-3858-8_389

Oura, G. K. (2001). Authentic task-based materials: Bringing the real world into the classroom. *Sophia Junior College Faculty Bulletin*, *21*, 65-84.

Oviedo, H. C., & Campo-Arias, A. (2005). Aproximación al uso del coeficiente alfa de Cronbach. *Revista colombiana de psiquiatría, 34*(4), 572-580.

Picaroni, B., & Loureiro, G. (2010). Qué matemáticas se enseña en aulas de sexto año de primaria en escuelas de Latinoamérica [What mathematics is taught in sixth grade classrooms in Latin American schools]. *Páginas de Educación, 3*(1), 29-60. https://doi.org/10.22235/pe.v3i1.657

Rodríguez-Martínez, L.Y. (2018). *Diseño y validación de un protocolo de observación para evaluar las actividades de enseñanza en quinto grado de primaria en la asignatura de Matemáticas*. [Tesis de Doctorado, Universidad Autónoma de Aguascalientes]. Repositorio bibliográfico Universidad Autónoma de Aguascalientes. http://hdl.handle.net/11317/1505

Schlesinger, L., & Jentsch, A. (2016). Theoretical and methodological challenges in measuring instructional quality in mathematics education using classroom observations. *ZDM Mathematics Education, 48*(1-2), 29-40. https://doi.org/10.1007/s11858-016-0765-0

Secretaría de Educación Pública. (2011). *Plan de Estudios. Educación Básica*. SEP.

Secretaría de Educación Pública. (2012). *Desafíos matemáticos [Mathematical Challenges]. Libro para el maestro. Quinto grado de primaria*. SEP.

Shavelson, R. J., & Webb, N. M. (1991). *A primer on generalizability theory*. Sage Publications.

Stein, M., Grover, B., & Henningsen, M. (1996). Building student capacity for mathematical thinking and reasoning: An analysis of mathematical task used in reform classroom. *American Educational Research Journal, 33*, 455-488. https://doi.org/10.3102/00028312033002455

Stiggins, R. (2007). Assessment for learning: An essential foundation of productive instruction. In R. Douglas (Ed.), *Ahead of the curve: The power of assessment to transform teaching and learning* (pp. 59-76). Solution Tree Press.

Stone, R. H., Boon, R. T., Fore III, C., Bender, W. N., & Spencer, V. G. (2008). Use of text maps to improve the reading comprehension skills among students in high school with emotional and behavioral disorders. *Behavioral Disorders*, *33*(2), 87-98. https://doi.org/10.1177/019874290803300203

Stylianides, A. J., & Stylianides, G. J. (2013). Seeking research-grounded solutions to problems of practice: Classroom-based interventions in mathematics education. *ZDM, 45*, 333-341. https://doi.org/10.1007/s11858-013-0501-y

Swan, M. (2015). *Designing tasks and lesson that develop conceptual understanding, strategic competence and critical awareness.* Center for Research in Mathematics Education: University of Nottingham.

Turner, R. C., Keiffer, E. A., & Salamo, G. J. (2018). Observing inquiry-based learning environments using the Scholastic Inquiry Observation Instrument. *International Journal of Science and Mathematics Education, 16*(1), 1455-1478. https://doi.org/10.1007/s10763-017-9843-1

UNESCO (2021). *Los aprendizajes fundamentales en America Latina y el Caribe. Estudio Regional Comparativo y Explicativo (ERCE 2019). Resumen ejecutivo*. UNESCO. https://unesdoc.unesco.org/ark:/48223/pf0000380257

Walkowiak, T. A., Berry, R. Q., Meyer, J. P., Rimm-Kaufman, S. E., & Ottmar, E. R. (2013). 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

Wiggins, G. (1998). *Educative assessment: Designing assessments to inform and improve student performance*. Jossey-Bass Publishers. https://doi.org/10.22235/pe.v1i1.718

Zolkower, B., & Bressan, A. (2012). Educacio□n Matema□tica Realista [Realistic Math Education]. In M. Pochulu & M. Rodri□guez (Eds.), *Educacio□n matema□tica. Aportes a la formacio□n docente desde distintos enfoques teo□ricos*. (pp. 175-200). Universitaria de Villa Mari□a y Universidad Nacional de Gral. Sarmiento.