European Journal of Science and Mathematics Education, 2023, **11**(2), 344-359
ISSN: 2301-251X (Online)

OPEN ACCESS

Research Article

Rasch analysis and validity of the construct understanding of the nature of models in Spanish-speaking students

Jose M. Oliva [1*]
0000-0002-2686-6131

Ángel Blanco [2]
0000-0003-3628-0801

[1] Department of Didactics, University of Cádiz, Cádiz, SPAIN
[2] Department of Science Education, University of Málaga, Málaga, SPAIN
[*] Corresponding author: josemaria.oliva@uca.es

Citation: Oliva, J. M., & Blanco, A. (2023). Rasch analysis and validity of the construct understanding of the nature of models in Spanish-speaking students. *European Journal of Science and Mathematics Education, 11*(2), 344-359. https://doi.org/10.30935/scimath/12651

ARTICLE INFO

Received: 23 Jun 2022

Accepted: 13 Nov 2022

ABSTRACT

A questionnaire was recently developed for the use with the Spanish-speaking, and evidence have been provided about the construct internal validity by means of structural equation modelling. In this paper, two research questions were considered: (i) What new evidence does application of the Rasch model provide regarding the validity of this construct? (ii) What cutoffs should be applied to the constructed scales in order to differentiate between acceptable and insufficient levels of the construct being measured? Participants were 1,272 Spanish at both high-school and college level. The instrument is a pencil and paper questionnaire written in Spanish, comprising 20 items (5-point Likert-type scale) distributed evenly across four scales: beyond exact replicas, purpose of models, multiple models, and changing models. Students' responses were coded on an ordinal scale from zero to four. We then conducted a Rasch analysis using both a multidimensional approach and a consecutive unidimensional approach for each dimension. Data provided new evidence regarding the internal validity of the four scales of the questionnaire. The Rasch analysis also allowed us to establish cutoffs for the constructed scales. The evidence provided by this, and the previous study suggest that the questionnaire may be useful as a diagnostic tool when applied to groups or populations of students. In addition, the identified cutoffs could, hypothetically, serve to differentiate between students with an adequate versus an insufficient understanding of the nature of models.

Keywords: cutoff point, nature of models, construct validity, Rasch model

INTRODUCTION

Models are central to both the practice and learning of science (Figueiredo & Perticarrari, 2022; Gilbert et al, 2000; Halloun, 1996; Justi, 2006; Karampelas, 2021; Martinand, 1986); hence, it is important that students learn about and are able to apply them. In order to do so, they must have an adequate understanding of the nature of models in science. This type of understanding has been referred to as epistemic knowledge, metamodeling knowledge, or metarepresentational competence, among other terms (Nicolaou & Constantinou, 2014; Schwarz, 2002), and its importance for science education is supported at both the theoretical and empirical level (Gobert et al., 2011; Schwarz, 1998, 2002; Schwarz & White, 2005; Sins et al., 2009). It is essential, therefore, to have suitable tools for assessing this construct.

Numerous authors have sought to characterize this kind of epistemic knowledge and to propose instruments for assessing it (Krell & Krüger, 2015; Krell et al., 2015; Treagust et al., 2002). One of the most widely known tools is the students' understanding of models in science (SUMS) (Treagust et al., 2002), which has been used in various studies (Cheng & Lin, 2015; Everett et al., 2009; Park et al., 2017; Treagust et al., 2002). It should be noted, however, that the results of a recent study suggest that the SUMS may not be adequate for assessing undergraduate science students' understanding of models and modeling (Lazenby & Becker, 2021).

With regard to instruments in Spanish, some authors have used closed questionnaires to assess students' understanding of the nature of models as part of a broader evaluation of their ideas about the nature of science and scientific knowledge in general (Pardo et al., 2018; Raviolo et al., 2010; Vasques Brandão et al., 2015), while others have used a qualitative approach to examine prospective teachers' understanding of models in science (Bravo & Mateo, 2017). More recently, a Spanish adaptation of the SUMS was tested and validated in a sample of Chilean students in grades 7 to 10 (Villablanca et al., 2020). Importantly, however, research conducted with Spanish-speaking students has also brought to light a number of difficulties with the translation and adaptation of many SUMS items (Oliva & Blanco-López, 2021). Thus, tests and questionnaires are particularly sensitive to cultural and linguistic aspects, and consequently their use in a context other than the one in which they were developed implies a complex process of adaptation. Although every language is translatable (Hunt & Agnoli, 1991), this often requires substantial changes since a natural expression in one language may by unacceptable in another. Therefore, each language has its own culturally-based rhetorical styles that imply different ways of organizing ideas (Connor, 1996). For example, the expression of ideas in English is normally much more direct and concise than it is in Romance languages such as Spanish (Kaplan & Grabe, 1991), and these stylistic aspects must also be adequately addressed when translating and adapting a measurement instrument.

With the aim of overcoming these limitations, Oliva & Blanco-López (2021) recently developed a new tool, the CoNaMo (based on its name in Spanish: Comprensión de la naturaleza de los modelos; in English: Understanding of the nature of models), which is inspired by the SUMS and designed to measure the same construct. The advantage of the CoNaMo is that it was developed directly for Spanish-speaking students, and its authors provide evidence supporting its application to students across a wide range of ages and levels of education (Oliva & Blanco-López, 2021). However, they also highlight the need for additional kinds of validity evidence, including a more in-depth examination of the response processes students follow when answering its items. One way of addressing this would be through the use of procedures based on item response theory (IRT) (Muñiz, 2010).

Our aim here, therefore, is to conduct a Rasch analysis (Rasch, 1960) of results obtained with the CoNaMo in order to provide further support for the construct it purports to measure, Understanding of the nature of models, thus complementing a previous study that analyzed its content validity based on expert opinion and the internal validity of scale scores by means of factor analysis. We consider that this new study could also provide additional information regarding the CoNaMo scales, such as possible cutoffs that could be used to discriminate between different levels of understanding of the nature of models.

THEORETICAL FRAMEWORK

Limitations of Classical Validation Methods

Classical test theory (CTT), which is undoubtedly the predominant framework for test development and analysis, assumes that an individual's level on a given construct, as measured by an assessment scale, can be determined through the linear combination of the partial results derived from the items used to evaluate this construct (Muñiz, 2010). However, the literature has shown that despite the utility of CTT, the procedures derived from it, such as factor analysis, have several limitations. First, although factor analysis is a powerful technique for investigating multidimensionality in observational data, it cannot construct interval measures from ordinal data, such as those obtained through Likert-type scales (Boone & Scantlebury, 2011; Neumann et al., 2011). Second, test results depend on the sample characteristics and the difficulty of the items chosen, and hence the instrument needs to be re-validated each time the participants change, and the results from different instruments are not comparable (Bond & Fox, 2007). Third, these procedures tell us little about the metrics used in the item response scales, in this case based on Likert-type scales. For example, how appropriate is it to evaluate a given construct with a five-category Likert-type scale?

These limitations can be overcome by other types of analysis such as those based on IRT, which are able to analyze aspects such as the appropriate number of response categories or the discriminatory power of a measurement scale. IRT-based methods, and in particular the Rasch model (Rasch, 1960), may be considered as complementary to traditional factor analytic methods, and in fact the two have been used in conjunction to analyze the psychometric properties of the scales that comprise the SUMS (Wei et al., 2014) and to compare students' understanding of models and modeling in relation to different scientific disciplines (Krell et al., 2015).

A further issue to consider concerns the identification of performance thresholds on a measurement scale that are able to differentiate between levels of the construct being measured, such that the results may be treated as qualitative categories. This is not a straightforward decision, although it is not advisable to choose as the reference threshold the central value of a Likert-type item response scale because scores on these scales are prone to inflation. In fact, some authors have criticized the fact that the vast majority of research on epistemic beliefs has used self-report surveys (Lindfors et al., 2020). It is argued that this type of instrument, particularly when consisting of closed items, may allow students to give "high-level" answers based on hearsay only (Sins et al., 2009), whereas qualitative studies suggest that students in fact have a naïve and unsophisticated epistemological understanding of models (Lazenby et al., 2019). However, although it is true that Likert-type questionnaires, may produce an inflated view of students' epistemic knowledge, it is also possible that a student does have adequate knowledge in this respect but is unable to demonstrate it when answering an open written question (Lazenby et al., 2019). This means that the data obtained with Likert-type self-report surveys should only be interpreted in relative terms, comparing results rather than treating them in absolute terms, or alternatively establishing thresholds or cutoffs that allow evaluators or teachers to differentiate between performance levels.

Various psychometric studies have described procedures for establishing cutoffs (Goodwin, 1996), many of which involve the use of an expert panel to determine the difficulty of each test question; having thus determined the minimum ability a candidate requires to get each question right, these expert ratings can then be compared with respondents' actual performance on the test (Angoff, 1971). In some cases, this comparison is supported by a logistic regression model, such as the Rasch model (Baghaei, 2007), and this further justifies the use of this type of analysis for providing new validity evidence about the construct measured by the CoNaMo.

Rasch Model

Rasch analysis (Rasch, 1960) is one of a group of mathematical models for test validation that are based on IRT, and which provide an alternative to the classical approach. It is a method for determining the psychometric properties of a set of items or questions, and it allows decisions to be made about the composition and structure of a questionnaire; it can also be used to create interval measures for latent scales (Liu & Boone, 2006). In contrast to classical statistical analyses, in which what is analyzed are the observed data, the mathematical model underlying Rasch analysis is derived from a logistic function in which the probability (P_{ni}) of a given response by individual *n* to item *i* is calculated based on the difference between the individual's ability level (θ_{n}) and the difficulty of the item (δ_i).

$$P_{ni} = e^{(\theta_n - \delta_i)} / \left(1 + e^{(\theta_n - \delta_i)}\right).$$

One advantage of the Rasch model is that it is applicable to ordinal data, and hence it is highly suited to the analysis of instruments based on Likert-type scales (Boone et al., 2010; Neumann et al., 2011). A further advantage is that it allows the performance of a group of students to be evaluated, irrespective of the difficulty of the items used or the ability level of respondents (Rasch, 1977). Finally, the Rasch model locates the person's performance and the item difficulty on the same latent variable, thus enabling direct comparison of these two parameters (Andrich & Marais, 2019). The values of latent variables may vary between minus infinity and plus infinity, and they are expressed in *logits*.

The key assumption of the Rasch model is that there is a functional relationship between the measure corresponding to a given item response category and the probability that a student chooses that response category. This function is known as the item characteristic curve. In polytomous scales, each item has several curves, one for each of its response categories, which are expected to follow a logical order in terms of their difficulty level (Masters, 1982). Any variation in this order, or the observation that one of the curves is subsumed by another, is indicative of problems in the metrics used, which can be resolved by reformulating one or more of the initial categories.

In order to analyze items, the Rasch model first transforms ordinal data into interval scales by applying a logistic function. It can then be used to evaluate different aspects such as the uni-dimensionality of measures, the degree to which the data fit the model, item difficulty, or person and item reliabilities and the corresponding separation indices (Boone et al., 2010; Liu & Boone, 2006). In the case of polytomous items, it is necessary to check whether the difficulty levels of item response categories are monotonically ordered. The aim here is to verify the quality of measures, that is to say, whether the data meet the specifications of a Rasch model.

Previous Validity Evidence for the CoNaMo

The CoNaMo (Oliva & Blanco-López, 2021) is one of a number of instruments that, like the SUMS, aims to assess students' understanding of the nature of models and their uses in science, in general rather than linked to specific contexts. Thus, in the tradition of many studies on the nature of science, the underlying premise here is that it is possible to establish certain characteristics that are common to models across different scientific disciplines (Van Der Valk et al., 2007). It is these common characteristics of models in science that are the focus of the CoNaMo. Of course, this does not mean there is no need to develop questionnaires aimed at evaluating students' understanding of models and modeling in specific contexts, as some authors have done (Gogolin & Krüger, 2018; Krell et al., 2014; Lee et al., 2015; Treagust et al., 2004). In our view, both perspectives (the general and the specific) are necessary as they complement one another.

Based on a theoretical study (Crawford & Cullin, 2005; Grosslight et al., 1991; Schwarz & White, 2005; Sins et al., 2009; Treagust et al., 2002) and on the results of previous studies that piloted Spanish versions of the SUMS (Jiménez-Tenorio et al., 2016; Muñoz-Campos et al., 2016), we developed a set of 50 items, ten for each of the five dimensions: *kinds of models*, *beyond exact replicas*, *purpose of models*, *multiple models*, and *changing models*. In order to assess the content validity of this first version of the instrument we asked a panel of 20 experts to rate the adequacy y clarity of each item. Based on the experts' feedback, we eliminated seven items, and we reformulated a further seven items whose wording was considered to be confusing by three or more experts. As a result of this process, we then created a closed questionnaire in which the 43 items were grouped according to their corresponding dimension, although within each they were randomly distributed. The response format for each item was a five-point Likert scale: strongly agree, agree, neither agree nor disagree, disagree, and strongly disagree.

Subsequently, we examined the internal structure of the resulting instrument by administering it to a wide sample of students aged between 14 and 55 years. Questionnaire responses were coded on an ordinal five level scale. Next we calculated descriptive statistics, item by item, in order to analyze item homogeneity. This enabled us to select those items that appeared to be most consistent with their corresponding dimension. After purifying the instrument based on the results of these analyses we obtained a final version of the questionnaire comprising 20 items distributed equally across four scales: *beyond exact replicas*, *purpose of models*, *multiple models*, and *changing models* (**Table 1**), (see **Appendix A**). The fifth dimension–*kinds of models*–was eliminated from the questionnaire, as it does not show adequate construct validity, and thus it is not possible to define a specific subscale for it.

We then performed a principal components analysis in order to study the dimensionality of the questionnaire, thus enabling us to group the 20 items by factors. Model fit was assessed by means of confirmatory factor analysis (CFA) and through structural equation modeling (SEM) (Oliva & Blanco-López, 2021). Moreover, he fact that the four scales were found to be inter-related suggests there is a second order latent factor that explains an important part of the variance in the four latent variables that were considered separately (Oliva & Blanco-López, 2021; Treagust et al., 2002). This construct, which we referred to as meta-modeling knowledge, may be regarded as testing a general factor reflecting students' overall understanding of the nature of models in science.

In light of the above and having provided evidence about the factorial validity of the construct measured by the CoNaMo (Oliva & Blanco-López, 2021), we consider it important to add further validity evidence based on the Rasch model. It would also be useful to analyze the metric of the five-level Likert scales used and to

Table 1. Dimensions of the concept of models and their uses in science, as distinguished in the CoNaMo

Dimension	Description	Authors
Beyond exact replicas	Models are a simplified or idealized representation of the real world, rather than an exact replica of it. Their focus is on aspects of the represented object, and other aspects will be ignored or considered as being of secondary importance, including interpretive components.	Jansen et al. (2019) Krell et al. (2015) Treagust et al. (2002)
Purpose of models	A model explains and predicts natural phenomena as well as to communicate scientific ideas and to test the theories related to it, insofar as models allow theories to be linked to the real world.	Justi and Gilbert (2003) Krell and Krüger (2015, 2017) Oh and Oh (2011) Treagust et al. (2002)
Multiple models	There are often multiple models of the same phenomenon because a variety of ideas can emerge about it, and because there may be different ways of representing it.	Crawford and Cullin (2004, 2005) Jansen et al. (2019) Oh and Oh (2011) Treagust et al. (2002)
Changing models	All models are open to revision and may, over time, change or ultimately be replaced with another.	Crawford and Cullin (2004, 2005) Jansen et al. (2019) Krell et al. (2015) Treagust et al. (2002)

determine possible cutoffs that could be used to discriminate between different levels of performance, thus converting the scales into qualitative variables. Indeed, this is a prerequisite to employing the CoNaMo scales for diagnostic or decision-making purposes, whether applied to individuals or groups of students. In this respect, it should be noted that Rasch analysis is increasingly being used in research examining students' difficulties and learning progressions in science (Neumann et al., 2011; Osborne et al., 2016; Rodríguez-Mora et al., 2022; Romine et al., 2020; Testa et al., 2019).

Research Questions

The purpose of the present study is therefore to address the following two questions:

1. What new evidence does application of the Rasch model provide regarding the validity of the construct Understanding of the nature of models and their uses in science, as measured by the CoNaMo?
2. What cutoffs should be applied to the scales that comprise the CoNaMo in order to differentiate between acceptable and insufficient levels of the construct being measured?

METHOD

Participants

For this study we used a sample comprising 1,272 Andalusian students (South of Spain). 580 were women (54.4%) and 692 were men (45.6% male), aged between 14 and 55 years (M=19.2; SD=5.3). One group of participants were students currently enrolled at one of five schools (three publics and two privates) in either compulsory secondary education (grades 10 and 11; 31.5%) or baccalaureate studies (grades 12 and 13; 18.2%). The remaining participants were adults from two public universities studying for a bachelor's or master's degree in either a science subject (chemistry, chemical engineering, biochemistry enology, and biotechnology; 22.2%) or teacher training (28.1%). The choice of participants in the sample was for convenience. However, since the number of subjects participating in the study was quite large, and the chosen schools and universities were diverse and representative of most Spanish educational institutions, it can be considered that the sample used was adequate for the purposes of the study.

The data collection process was carried out in accordance with current Spanish legislation regarding the protection of students' data and all participating students gave informed consent to participate in the study.

Instrument

The CoNaMo is a pencil-and-paper questionnaire comprising 20 items distributed evenly across four scales: beyond exact replicas (items 1 to 5), purpose of models (items 6 to 10), multiple models (items 11 to 15), and changing models (items 16 al 20). Half of the items are positively worded with respect to an epistemologically adequate view, while the remainder are reversed; each scale contains at least two positive and two reverse worded items. Respondents rate each item on a 5-point Likert-type scale, with the following options: strongly agree, agree, neither agree nor disagree, disagree, strongly disagree (see **Appendix A**).

The previous validation study of the CoNaMo showed that the questionnaire was applicable to a wide range of ages and levels of education, from 14-year-old students through to science undergraduates and prospective elementary and secondary science teachers (Oliva & Blanco-López, 2021). No time limit was set for completing the questionnaire. Instructions for instrument administrators are included in **Appendix B**.

Procedure and Data Analysis

Students' responses to the CoNaMo were coded on an ordinal scale from zero (corresponding to strongly disagree) to four (strongly agree). Scores on reverse worded items were then recoded, inverting the values. Higher scores therefore indicated a more advanced understanding of the nature of models. In order to reduce the amount of information, we summed the item scores for each dimension of the CoNaMo, thus obtaining a scale ranging from 0 to 20, or from 0 to 1 if the values are normalized to 1. We will refer these scores as *raw scores*. More details for questionnaire rating are included in **Appendix C**.

We then conducted a Rasch analysis using the partial credit model, which is the most suitable for Likert-type scales (Masters, 1982). Drawing on Briggs and Wilson's (2003) discussion of different ways of dealing with multiple dimensions, we opted to use a multidimensional approach, which is the most appropriate in the case of a questionnaire, such as the CoNaMo, that has four distinct dimensions grouped under a second-order factor. The software used for this analysis was *Constructmap 4.6* (Kennedy et al., 2010), which allowed us to obtain latent scales of scores, which we will call *measures*, one for each dimension, and which always fell within the range -3 to +3. In parallel, however, we also analyzed the data using a consecutive unidimensional approach, in which each dimension is considered separately and modeled independently as a unidimensional construct. For this analysis we used *Winsteps 4.4.7* (Linacre, 2020). Although the consecutive approach is less suited to the study of multidimensional scales, it is more parsimonious (Briggs & Wilson, 2003) and provides a more complete analysis of indicators for each of the separate scales.

Regarding the identification of cutoffs for the CoNaMo scales, the usual procedure with Rasch analysis is, as noted earlier, to establish thresholds based on the ratings of experts and then to calibrate these with respect to students' actual performance, before locating them on the latent scales (Baghaei, 2007). For the present study, however, we will use a different approach that we have not found described in the literature, and which may therefore be considered novel. Specifically, we will analyze directly the measures obtained with the latent scales constructed through the Rasch analysis, the aim being to identify inductively possible discontinuities between item response categories. These discontinuities could suggest meaningful cutoff points along the latent scales (*measures*), indicating a qualitative leap or advance that should be considered. To this end we compute Rasch-Thurstone thresholds, as recommended when the aim is to dichotomize a sample (Linacre, 2009). It should be noted that whereas the *measures* are expressed as linear or interval values, the *raw scores* are ordinal and hence cannot be used as the basis for identifying discontinuities or qualitative leaps directly.

RESULTS

We will begin by describing the results obtained regarding uni-dimensionality and the fit and reliability of the scales constructed through the Rasch analysis, as well as the item category probability curves and Wright maps resulting from the analyses conducted. These results will provide evidence for the internal validity of the measured construct (research question 1). The Wright maps will also be used with the aim of identifying possible discontinuities in the constructed scales (research question 2).

Uni-Dimensionality of Measures

For the Rasch model to yield precise estimates the constructed scales must be unidimensional. In the case of the CoNaMo, the four scales were shown to fulfill this condition in the previous study that used SEM (Oliva & Blanco-López, 2021). Our aim here was to corroborate this by applying principal components analysis of standardized residuals to each dimension. This analysis indicated an explained variance of 50%, 43%, 40%,

Table 2. Fit indicators for the 20 items of the CoNaMo

Dimension	Item	Multidimensional Rasch			Consecutive unidimensional Rasch		
		Measure (logit)	Infit MNSQ	Outfit MNSQ	Measure (logit)	Infit MNSQ	Outfit MNSQ
Beyond exact replicas	1	.09	1.11	1.14	.10	1.18	1.16
	2	.03	.97	.99	.08	.84	.79
	3	.26	.98	.99	.47	.81	.78
	4	-.17	.90	.93	-.30	.96	.95
	5	-.20	1.01	.99	-.35	1.17	1.16
Purpose of models	6	.17	.90	.90	.25	1.07	1.12
	7	-.41	.94	.94	-.48	1.00	1.10
	8	.25	.87	.87	.25	1.01	1.03
	9	-.02	.76	.76	-.08	.88	.82
	10	.01	1.15	1.15	.05	.95	.96
Multiple models	11	.10	.78	.77	.05	1.06	1.05
	12	-.28	.79	.81	-.44	.89	.88
	13	.04	.75	.74	.02	1.01	1.01
	14	.14	1.00	.99	.35	1.18	1.25
	15	.00	1.20	1.21	.02	.84	.83
Changing models	16	.08	.82	.90	.05	1.27	1.29
	17	-.01	.84	.91	.19	.95	.99
	18	-.08	.61	.67	.00	.78	.80
	19	.12	.66	.75	.09	.97	.96
	20	-.11	1.36	1.25	-.34	1.08	.97

Table 3. Separation and reliability indices for the four dimensions of the CoNaMo

Dimension	Separation		Reliability	
	Persons (students)	Items	Persons (students)	Items
Beyond exact replicas	1.65	7.13	.73	.98
Purpose of models	1.41	6.59	.67	.98
Multiple models	1.36	6.08	.65	.97
Changing models	1.48	3.86	.68	.98

and 43%, respectively. The corresponding values for the percentage of unexplained variance in the first contrast were 1.55, 1.52, 1.44, and 1.46 (i.e., less than 2 in all cases). These data suggest sufficient unidimensionality for the subscales considered.

Fit Indicators

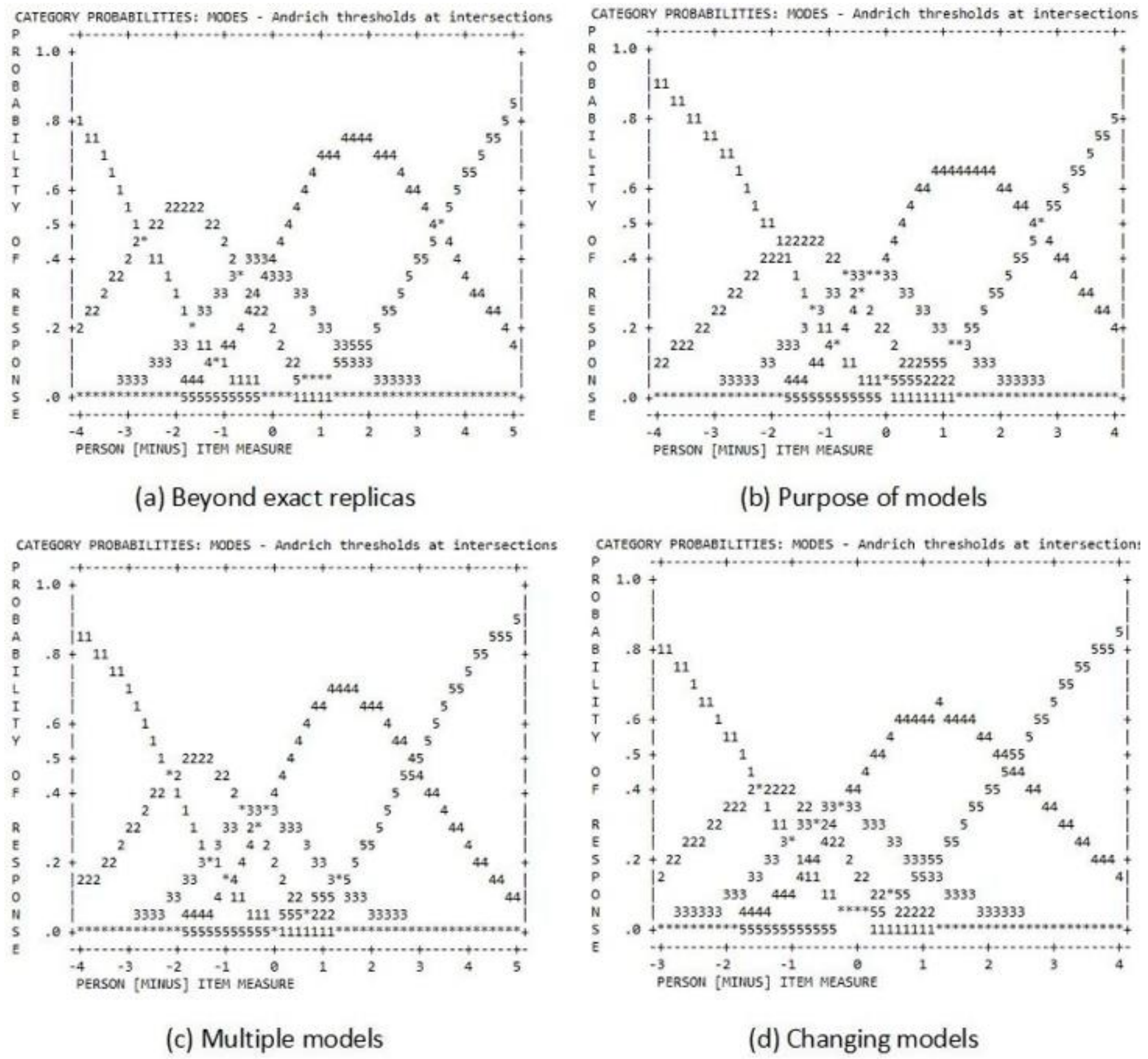
Table 2 shows the item measures (in logits) obtained in both the multidimensional analysis and the consecutive unidimensional analysis. Here, each item is located along a difficulty continuum constructed for each dimension, such that higher and lower values indicate more complex and simpler items, respectively.

Also displayed in **Table 2** are the infit and outfit mean square errors (MNSQ) for items in both analyses. Values close to one are desirable here, whereas very small values imply local dependence of items and high values indicate randomness in the data. The results in this case indicated good fit of the data to both the multidimensional and the consecutive Rasch model. Specifically, infit values in the multidimensional model ranged from .61 to 1.36, while outfit values were between .67 and 1.25. For the consecutive unidimensional model, infit ranged from .78 to 1.27, while outfit values were between .78 and 1.29. All these values fall within the recommended range of 0.60-1.40 (Wright & Linacre, 1994). Note also that the results obtained are similar for the two approaches.

Reliability and Separation Indices

With the consecutive unidimensional approach, *Winsteps* computes reliability and separation indices for the constructed scales. Specifically, for both reliability and separation it yields two distinct indicators, one for persons (in this study, students) and one for items. These indicators are measures of scale quality, insofar as they reflect the ability of model estimates to discriminate between different levels of the measures.

Table 3 shows the separation and reliability indices for each scale of the CoNaMo. Values of both indices were very high for items, while for persons they were lower but sufficient. In fact, the mean value of the person



(a) Beyond exact replicas

(b) Purpose of models

(c) Multiple models

(d) Changing models

Figure 1. Probability curves for items of the different dimensions (difficulty increases from left to right on the logit scale) (Source: Authors)

separation index was close to 1.5 (equivalent to a reliability of .7), which is the reference value regarded as acceptable (Fisher, 1992).

Probability Curves

Winsteps can also be used to plot item category probability curves for each dimension of a measurement instrument. This is a useful way of testing:

(1) if all the categories of the Likert-type scale used are relevant or whether one or more of them is subsumed by another, in which case they would need to be reformulated, and

(2) if the logit measures of the item categories are ordered as expected.

Figure 1 shows the probability curves for the four scales of the CoNaMo.

It can be seen that for each scale the curves for the different categories follow the expected order in terms of difficulty: level 4 is more difficult than level 3, which in turn is more difficult than level 2, and so forth to level 0. In addition, none of the categories is completely subsumed by another, even though the curves for levels 1 and 2, and especially the latter, show considerable overlap with those of the adjacent categories.

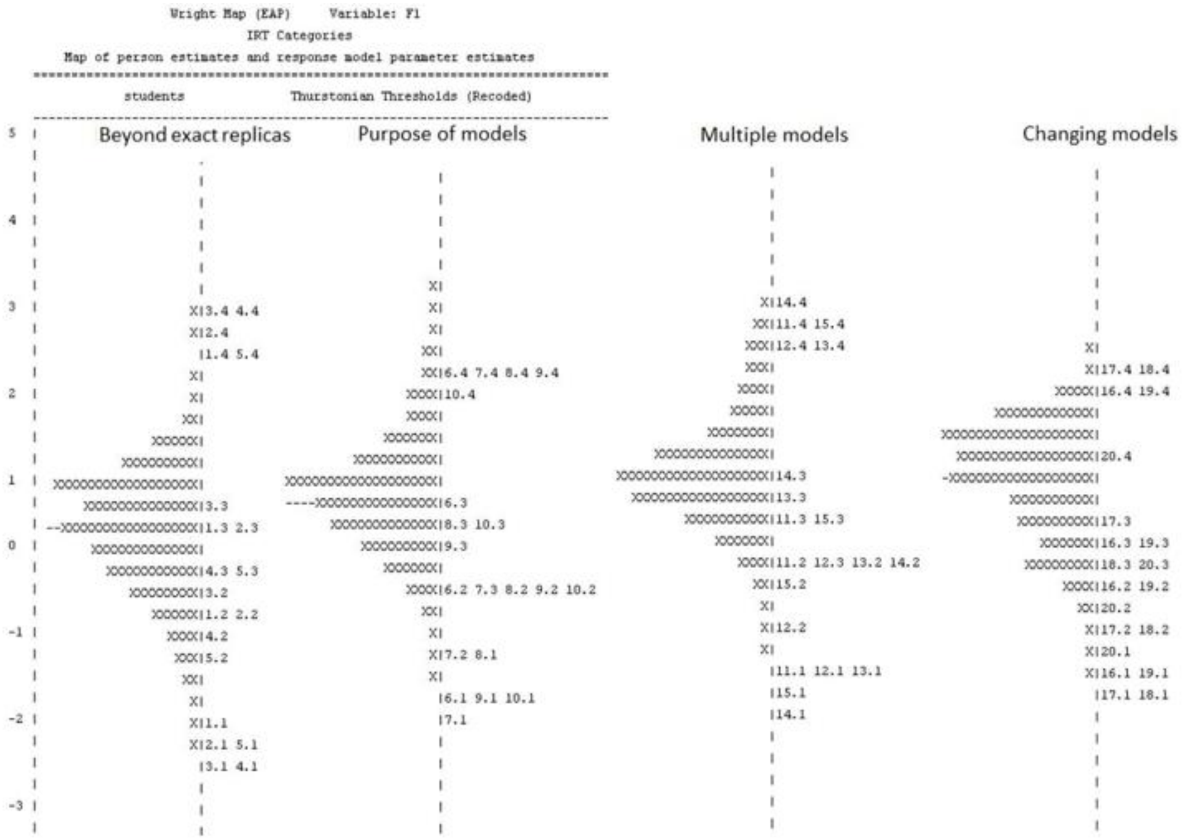


Figure 2. Wright map showing the spread of students' performance (person estimates) and the Thurstonian thresholds for items (Source: Authors)

Wright Maps

The Wright map is one of the most useful outputs of the Rasch model, insofar as it shows the person and item estimates located on the same latent scale. **Figure 2** displays the Wright maps for the four scales of the CoNaMo, in this case corresponding to the multidimensional Rasch model.

The letters "x" on the left side of each axis correspond to person estimates on the latent scale. Those whose parameter estimates are toward the upper part of the scale are more competent on the corresponding dimension, while those towards the bottom are less so. On the right side of each axis, the map shows the Thurstonian thresholds for the response categories of the different items. For example, the logit threshold value of one for item 3 (shown on the map as 3.1) expresses the value of the latent variable Beyond exact replicas, for which the probability that students choose level 1 or higher on the Likert-type scale is greater than the probability of their choosing level 0. These thresholds therefore represent measures of the transitions between categories. The response categories corresponding to Thurstonian thresholds toward the upper part of the scale are more demanding (difficult), whereas those towards the bottom are easier.

Maps of this kind allow us to visualize whether or not the different categories or levels of item measures cover the range of the distribution of person (student) measures, something which is always desirable in a questionnaire. It can be seen that for the sample analyzed here there is high variability in measures on the four scales, both for persons and item categories. Ideally, one would observe maximum possible overlap between the two parameters, as this would indicate that the scales used are highly suitable for measuring students' performance. This is indeed the case here (**Figure 2**) since the item category and person parameters show targeting.

However, there are also gaps in the spread of categories across the four scales. Whereas the Thurstonian thresholds for response categories 1, 2, and 3 on the four scales are close to and show continuity with each other, there is an important gap between these three categories and level 4, especially on the scales beyond exact replicas, purpose of models, and multiple models. The gap is also observable for the other scale, changing models, although in this case, category 4 of item 20 is located here. These gaps, which on the latent scales appear around a value of 1.0 logits, reveal qualitative leaps or discontinuities in the latent scale values of item response categories, specifically between categories 3 and 4 on the Likert-type scale. This, as we will discuss below, is useful for identifying cutoff points.

DISCUSSION

Our first goal in this study was to provide new validity evidence for the construct measured by the CoNaMo (i.e., *Understanding of the nature of models in science*), adding to that reported in a previous study which used factor analysis and SEM (Oliva & Blanco-López, 2021). It is important to remember that procedures based on factor analysis have important limitations (Bond & Fox, 2007), especially when an instrument uses ordinal Likert-type scales, which is usually the case in studies related to the nature of science (Neumann et al., 2011). As we have seen here, the four scales that comprise the CoNaMo show a good fit to the Rasch model. Item reliability indices were also high, although person indices were only acceptable. This means that while the CoNaMo scales are sufficiently reliable to be used to compare item measures, or to compare the performance of groups of students, they are less suited to decision making based on individual student measures. An increase in person separation reliability would probably require the addition of new items to each dimension, although this would then mean the inclusion of reiterative or redundant statements; this, in turn, would likely produce local dependence between items, and hence a poorer fit of the data to the model. We must therefore accept this limited reliability of person measures as a limitation of the constructed scales, one that they share with the SUMS scales (Wei et al., 2014).

Regarding the probability curves, these indicated acceptable behavior, with the five response categories being ordered monotonically, as expected. However, and as in the Rasch validation of the SUMS (Wei et al., 2014), some of the intermediate categories on the Likert-type scale used overlapped with the adjacent categories, suggesting the need to revise them.

The Wright maps provide useful visual information about the distributions of measures along the latent dimensions, and they allow comparisons to be made. As we have seen, there is good overlap or targeting between the person measures and the item Thurstonian thresholds, which is desirable in terms of scale quality. However, and related to our second research question, there are also discontinuities in the spread of threshold values that could help to establish cutoff points in the measures obtained. In particular, there is a discontinuity in the latent variable when moving from level 3 to level 4 on the Likert-type scale, that is, between the categories *agree* and *strongly agree* for positively worded items, and between *disagree* and *strongly disagree* for reversed items. This confirms the need to treat both the Likert-type scales of the CoNaMo and the raw scores obtained from them as ordinal rather than interval-level variables. In order to work with the latter, we would need to transform the raw scores into measures on the latent dimensions.

Another possibility would be to transform the quantitative raw scores into dichotomous nominal variables, distinguishing between adequate and insufficient levels of performance based on the score obtained. This could be achieved by establishing a cutoff based on the Thurstonian thresholds, which would locate it at category 3 on the Likert-type scale, or around a value of 1.0 logits on the latent scales. This value is equivalent to a score of 15 on a scale from 0 to 20, or to a value of .75 if the scale for each dimension is normalized to a range from 0 to 1. This qualitative reinterpretation could have interesting repercussions. Specifically, it could eliminate the inflation effect that usually contaminates the results obtained with Likert-type scales, thus avoiding discrepancies between the conclusions that emerge from studies which use this type of instrument and those which employ qualitative measures (Lazenby et al., 2019; Sins et al., 2009). Use of the aforementioned cutoff would mean, for example, that a raw score on the CoNaMo of 12 out of 20 (.6 on the scale zero to one) would be considered a low value, despite being above the midpoint of the scale.

CONCLUSIONS AND IMPLICATIONS

The results of this and a previous validation study that used factor analysis and SEM (Oliva & Blanco-López, 2021) show that students' understanding of the nature of models and their uses in science is a complex construct comprising several dimensions. Accordingly, the CoNaMo provides information about four different dimensions that, despite some degree of intercorrelation, are not sufficiently related to justify treating the measured construct as unidimensional. This is consistent with the results of other studies in this field, which usually define separate scales or dimensions for each of the sub-constructs they consider (Crawford & Cullin, 2005; Grünkorn et al., 2014; Jansen et al., 2019; Treagust et al., 2002).

Overall, the results of the Rasch analysis suggest that the CoNaMo may be useful as a diagnostic tool when applied to groups or populations of students, although it should not be used for diagnostic or assessment purposes with individual students, especially if this would have academic consequences. Our aim in the near future is to build on this research by using the CoNaMo as an assessment instrument in various studies: Analyzing changes in students' understanding of the nature of models in science across different levels of education; comparing the results obtained by college students from different science disciplines; studying the relationship between the results obtained by students and practicing science teachers; examining changes in questionnaire responses during teaching-learning sequences aimed at developing students' understanding of the nature of models; exploring the possible relationship between students' understanding of the nature of models and their ideas about, for example, the nature of science.

We consider that the results of this study contribute to further research in science education at least in the following aspects:

(1) they provide additional types of evidence of internal validity of the measured construct,

(2) they allow us to analyze the metrics of the 5-point Likert scales, and

(3) the Rasch analysis was useful to identify a cutoff point that could be used to transform the quantitative raw scores obtained with the CoNaMo scales into qualitative variables; specifically, it located the cutoff point at a score of 15 out of 20, or .75 if the scale is normalized to a range from zero to one.

This threshold could, hypothetically, serve to differentiate between students with an adequate versus an insufficient understanding of the nature of models and their uses in science. Testing this would require new complementary studies that combine quantitative and qualitative methods of data analysis, for example, comparing students' responses to interview questions or open-ended questionnaires with what is suggested by their results on the nominal CoNaMo scales, with the aforementioned cutoff.

Author contributions: All authors were involved in concept, design, collection of data, interpretation, writing, and critically revising the article. All authors approve final version of the article.

Funding: The authors received financial support from Spanish Ministry of Science, Innovation and Universities (Grant/Award Numbers: EDU2017-82518-P, EDU2017-82197).

Ethics declaration: Authors declared that at the time of this research was carried out, both the University of Cádiz and the University of Málaga had not developed any protocol for this type of study. Nevertheless, in the data collection, the Ethical Guidelines for Educational Research, fourth edition (2018), have been considered, respecting the following ethical principles in relation to the participating subjects: voluntary participation, informed consent, confidentiality, and privacy.

Declaration of interest: Authors declare no competing interest.

Data availability: Data generated or analyzed during this study are available from the authors on request.

REFERENCES

Andrich, D., & Marais, I. (2019). *A course in Rasch measurement theory: Measuring in the educational, social and health sciences*. Springer. https://doi.org/10.1007/978-981-13-7496-8

Angoff, W. H. (1971). *Scales, norms, and equivalent scores.* In R. L. Thorndike (Ed.), *Educational measurement* (pp. 508-600). American Council on Education.

Baghaei, P. (2007). Applying the Rasch rating-scale model to set multiple cut-offs. *Rasch Measurement Transactions, 20*(4), 1075-1076.

Bond, T. G., & Fox, C. M. (2007). *Applying the Rasch model: Fundamental measurement in the human science*. Lawrence Erlbaum Associates.

Boone, W. J., & Scantlebury, K. (2006). The role of Rasch analysis when conducting science education research utilizing multiple-choice tests. *Science Education, 90*(2), 253-269. https://doi.org/10.1002/sce.20413

Boone, W. J., Townsend, J. S., & Staver, J. (2010). Using Rasch theory to guide the practice of survey development and survey data analysis in science education and to inform science reform efforts: An exemplar utilizing STEBI self-efficacy data. *Science Education, 95*(2), 258-280. https://doi.org/10.1002/sce.20413

Bravo, B., & Mateo, E. (2017). Visión de los maestros en formación sobre los modelos científicos y sus funciones en las ciencias y en su enseñanza [Prospective teachers' understanding of scientific models and their role in science and science education]. *Didáctica de las Ciencias Experimentales y Sociales* [*Didactics of Experimental and Social Sciences*], *33*, 143-160. http://doi.org/10.7203/DCES.33.10102

Briggs, D. C., & Wilson, M. (2003). An introduction to multidimensional measurement using Rasch models. *Journal of Applied Measurement, 4*(1), 87-100.

Cheng, M. F., & Lin, J. L. (2015). Investigating the relationship between students' views of scientific models and their development of models. *International Journal of Science Education, 37*(15), 2453-2475. https://doi.org/10.1080/09500693.2015.1082671

Connor, U. (1996). *Contrastive rhetoric. Cross-cultural aspects of second-language writing*. Cambridge University Press. https://doi.org/10.1017/CBO9781139524599

Crawford, B., & Cullin, M. (2005). Dynamic assessments of pre-service teachers' knowledge of models and modelling. In K. Boersma, H. Eijkelhof, M. Goedhart, & O. Jong (Eds.), *Research and the quality of science education* (pp. 309-323). Springer. https://doi.org/10.1007/1-4020-3673-6_25

Everett, S. A., Otto, C. A., & Luera, G. L. (2009). Preservice elementary teachers' growth in knowledge of models in a science Capstone course. *International Journal of Science and Mathematics Education, 7*(6), 1201-1225. https://doi.org/10.1007/s10763-009-9158-y

Figueiredo, A. O., & Perticarrari, A. (2022). El aprendizaje basado en modelos mantiene a los alumnos activos y con atención sostenida [Model-based learning keeps learners active and focused]. *Revista Eureka sobre Enseñanza y Divulgación de las Ciencias, 19*(3), 3102. https://doi.org/10.25267/Rev_Eureka_ensen_divulg_cienc.2022.v19.i3.3102

Fisher, W. P. (1992). Reliability, separation, strata statistics. *Rasch Measurement Transactions, 6*(3), 238.

Gilbert, J., Boulter, C., & Elmer, R. (2000). Positioning models in science education and in design and technology education. In J. K. Gilbert, & C. J. Boulter (Eds), *Developing models in science education* (pp. 3-17). Kluwer. https://doi.org/10.1007/978-94-010-0876-1_1

Gobert, J., O'Dwyer, L., Horwitz, P., Buckley, B., Levy, S. T., & Wilensky, U. (2011). Examining the relationship between students' epistemologies of models and conceptual learning in three science domains: Biology, physics, & chemistry. *International Journal of Science Education, 33*(5), 653-684. https://doi.org/10.1080/09500691003720671

Gogolin, S., & Krüger, D. (2018). Students' understanding of the nature and purpose of models. *Journal of Research in Science Teaching, 55*(9), 1313-1338. https://doi.org/10.1002/tea.21453

Goodwin, L. D. (1996). Determining cut-off scores. *Research in Nursing & Health, 19*, 249-256. https://doi.org/10.1002/(SICI)1098-240X(199606)19:3<249::AID-NUR8>3.0.CO;2-K

Grosslight, L., Unger, C., Jay, E., & Smith, C. L. (1991). Understanding models and their use in science: Conceptions of middle and high school students and experts. *Journal of Research in Science Teaching, 28*(9), 799-822. http://doi.org/10.1002/tea.3660280907

Grünkorn, J., Upmeier zu Belzen, A., & Krüger, D. (2014). Assessing students' understandings of biological models and their use in science to evaluate a theoretical framework. *International Journal of Science Education, 36*(10), 1651-1684, https://doi.org/10.1080/09500693.2013.873155

Halloun, I. (1996). Schematic modelling for meaningful learning of physics. *Journal of Research in Science Teaching, 33*(9), 1019-1041. https://doi.org/10.1002/(SICI)1098-2736(199611)33:9<1019::AID-TEA4>3.0.CO;2-I

Hunt, E., & F. Agnoli (1991). The Whorfian hypothesis: A cognitive psychology perspective. *Psychological Review, 98*, 377-389. https://doi.org/10.1037/0033-295X.98.3.377

Jansen, S., Knippels, M. C. P. J., & van Joolingen, W. R. (2019). Assessing students' understanding of models of biological processes: A revised framework. *International Journal of Science Education, 41*(8), 981-994. http://doi.org/10.1080/09500693.2019.1582821

Jiménez-Tenorio, N., Aragón, L., Blanco, Á., & Oliva, J. M. (2016). Comprensión acerca de la naturaleza de los modelos por parte de profesorado de ciencias de secundaria en formación inicial [Understanding about the nature of models by preservice secondary science teachers]. *Campo Abierto, 35*(1), 121-132.

Justi, R. (2006). La enseñanza de ciencias basada en la elaboración de modelos [Teaching science through the development of models]. *Enseñanza de las Ciencias* [*Science Education*], *24*(2), 173-184. https://doi.org/10.5565/rev/ensciencias.3798

Justi, R. S., & Gilbert, J. K. (2003). Teachers' views on the nature of models. *International Journal of Science Education, 25*(11), 1369-1386. https://doi.org/10.1080/0950069032000070324

Kaplan, E., & Grabe, W. (1991). The fiction in science writing. In H. Schröder (Ed.), *Subject-oriented texts: Languages for special purposes and text theory* (pp. 199-217). Mouton de Gruyter. https://doi.org/10.1515/9783110858747.199

Karampelas, K. (2021). Trends on science education research topics in education journals. *European Journal of Science and Mathematics Education, 9*(1), 1-12. https://doi.org/10.30935/scimath/9556

Kennedy, C. A., Wilson, M. R., Draney, K., Tutunciyan, S., & Vorp, R. (2010). *ConstructMap 4.6*. BEAR Center.

Krell, M., & Krüger, D. (2015). Testing models: A key aspect to promote teaching activities related to models and modelling in biology lessons? *Journal of Biological Education, 50*(2), 160-173. https://doi.org/10.1080/00219266.2015.1028570

Krell, M., Reinisch, B., & Krüger, D. (2015). Analyzing students' understanding of models and modeling referring to the disciplines biology, chemistry, and physics. *Research in Science Education, 45*, 367-393. https://doi.org/10.1007/s11165-014-9427-9

Krell, M., Upmeier zu Belzen, A., & Krüger, D. (2014). Students' levels of understanding models and modelling in biology. *Research in Science Education, 44*, 109-132. https://doi.org/10.1007/s11165-013-9365-y

Lazenby, K., & Becker, N. M. (2021). Evaluation of the students' understanding of models in science (SUMS) for use in undergraduate chemistry. *Chemistry Education Research and Practice, 22*, 62-76. https://doi.org/10.1039/d0rp00084a

Lazenby, K., Stricker, A., Brandriet, A., Rupp, C. A., Mauger-Sonnek, K., & Becker, N. M. (2019). Mapping undergraduate chemistry students' epistemic ideas about models and modeling. *Journal of Research in Science Teaching, 57*(5), 794-824. https://doi.org/10.1002/tea.21614

Lee, S., Chang, H., & Wu, H. (2015). Students' views of scientific models and modeling: Do representational characteristics of models and students' educational levels matter? *Research in Science Education, 47*, 305-32. https://doi.org/10.1007/s11165-015-9502-x

Linacre, J. M. (2009). Dichotomizing rating scales and Rasch-Thurstone thresholds. *Rasch Measurement Transactions, 23*(3), 1228.

Linacre, J. M. (2020). *A user's guide to Winsteps/Ministeps Rasch model programs*. MESA Press.

Lindfors, M., Bodin M., & Simon, S. (2020). Unpacking students' epistemic cognition in a physics problem-solving environment. *Journal of Research in Science Teaching, 57*, 695-732. https://doi.org/10.1002/tea.21606

Liu, X., & Boone, W. J. (2006). Introduction to Rasch measurement in science education. In X. Liu, & W. J. Boone (Eds.), *Applications of Rasch measurement in science education* (pp. 1-22). JAM Press.

Martinand, J. L. (1986). Enseñanza y aprendizaje de la modelización [Teaching and learning modeling]. *Enseñanza de las Ciencias* [*Science Education*], *4*(1), 45-50. https://doi.org/10.5565/rev/ensciencias.5189

Masters, G. N. (1982). A Rasch model for partial credit scoring. *Psychometrika, 47*, 149-174. https://doi.org/10.1007/BF02296272

Muñiz, J. (2010). Las teorías de los tests: Teoría clásica y teoría de respuesta a los ítems [Test theory: Classical test theory and item response theory]. *Papeles del Psicólogo* [*Papers of the Psychologist*], *31*(1), 57-66

Muñoz-Campos, V., Cañero-Arias, J., Oliva-Martínez, J. Mª., Blanco-López, A., & Franco-Mariscal, A. J. (2016). Assessment of teacher training students' understanding of the nature of the models. In J. Lavonen, K. Juuti, J. Lampiselkä, A. Uitto, & K. Hahl (Eds.), *Science education research: Engaging learners for a sustainable future* (pp. 799-805). ESERA.

Neumann, I., Neumann, K., & Nehm, R. (2011). Evaluating instrument quality in science education: Rasch-based analyses of a nature of science test. *International Journal of Science Education, 33*(10), 1373-1405. https://doi.org/10.1080/09500693.2010.511297

Nicolaou, C. T., & Constantinou, C. P. (2014). Assessment of the modeling competence: A systematic review and synthesis of empirical research. *Educational Research Review, 13*, 52-73. https://doi.org/10.1016/j.edurev.2014.10.001

Oh, P. S., & Oh, S. J. (2011). What teachers of science need to know about models: An overview. *International Journal of Science Education, 33*(8,) 1109-1130. https://doi.org/10.1080/09500693.2010.502191

Oliva, J. M., & Blanco-López, A. (2021). Development of a questionnaire for assessing Spanish-speaking students' understanding of the nature of models and their uses in science. *Journal of Research in Science Teaching, 58*(6), 852-878. https://doi.org/10.1002/tea.21681

Osborne, J. F., Henderson, J. B., MacPherson, A., Szu, E., Wild, A., & Yao, S. Y. (2016). The development and validation of a learning progression for argumentation in science. *Journal of Research in Science Teaching, 53*(6), 821-846. https://doi.org/10.1002/tea.21316

Pardo, O, Solaz-Pórtoles, J. J., & San José, V. (2018). Creencias de los estudiantes de educación secundaria sobre la naturaleza de la ciencia y los modelos científicos: Un estudio transversal [Secondary education students' beliefs about the nature of science and scientific models: A cross-sectional study]. *Educatio Siglo XXI* [*21st Century Education*]*, 36*(3), 465-484. https://doi.org/10.6018/j/350091

Park, M., Liu, X., Smith, E., & Waight, N. (2017). The effect of computer models as formative assessment on student understanding of the nature of models. *Chemistry Education Research and Practice, 18*(4), 572-581. https://doi.org/10.1039/c7rp00018a

Rasch, G. (1960). *Probabilistic models for some attainment and intelligence tests.* Denmark's Pedagogical Institute.

Rasch, G. (1977). On specific objectivity: An attempt at formalizing the request for generality and validity of scientific statements. In M. Glegvad (Ed.), *The Danish yearbook of philosophy (pp. 59-94)*. Munksgarrd. https://doi.org/10.1163/24689300-01401006

Raviolo, A., Ramírez, P., López, E. A., & Aguilar, A. (2010). Concepciones sobre el conocimiento y los modelos científicos: Un estudio preliminar [Conceptions of knowledge and models in science: A preliminary study]. *Formación Universitaria* [*University Education*]*, 3*(5), 29-36.

Rodríguez-Mora, F., Cebrián-Robles, D., & Blanco-López, Á. (2022). An assessment using Rubrics and the Rasch Model of 14/15-year-old students' difficulties in arguing about bottled water consumption. *Research in Science Education, 52*, 1075-1091. https://doi.org/10.1007/s11165-020-09985-z

Romine, W. L., Sadler, T. D., Dauer, J. M., & Kinslow, A. (2020). Measurement of socio-scientific reasoning (SSR) and exploration of SSR as a progression of competencies. *International Journal of Science Education, 42*(18), 2981-3002. https://doi.org/10.1080/09500693.2020.1849853

Schwarz, C. V. (1998). *Developing students' understanding of scientific modelling* [Unpublished doctoral dissertation]. University of California, Berkeley.

Schwarz, C. V. (2002). Is there a connection? The role of meta-modeling knowledge in learning with models. In *Proceedings of the International Conference of Learning Sciences*.

Schwarz, C. V., & White, B. Y. (2005). Metamodeling knowledge: Developing students' understanding of scientific modeling. *Cognition and Instruction, 23*(2), 165-205. https://doi.org/10.1207/s1532690xci2302_1

Sins, P. H., Savelsbergh, E. R., van Joolingen, W. R., & van Hout-Wolters, B. H. (2009). The relation between students' epistemological understanding of computer models and their cognitive processing on a modelling task. *International Journal of Science Education, 31*(9), 1205-1229. https://doi.org/10.1080/09500690802192181

Testa, I., Capasso, G., Colantonio, A., Galano, S., Marzoli, I., di Uccio, U.S., Trani, F., & Zappia, A. (2019). Development and validation of a university students' progression in learning quantum mechanics through exploratory factor analysis and Rasch analysis. *International Journal of Science Education, 41*(3), 388-417. https://doi.org/10.1080/09500693.2018.1556414

Treagust, D. F., Chittleborough, G., & Mamiala, T. L. (2002). Students' understanding of the role of scientific models in learning science. *International Journal of Science Education, 24*(4), 357-368. https://doi.org/10.1080/09500690110066485

Treagust, D. F., Chittleborough, G., & Mamiala, T. L. (2004). Students' understanding of the descriptive and predictive nature of teaching models in organic chemistry. *Research in Science Education, 34*(1), 1-20. https://doi.org/10.1023/B:RISE.0000020885.41497.ed

Van Der Valk, T., Van Driel, J., & De Vos, W. (2007). Common characteristics of models in present-day scientific practice. *Research in Science Education, 37*(4), 469-488. https://doi.org/10.1007/s11165-006-9036-3

Vasques Brandão, R., Solano Araujo, I., & Veit, E. A. (2015). Validación de un cuestionario para investigar concepciones de profesores sobre ciencia y modelado científico en el contexto de la física [Validation of a questionnaire for investigating teachers' ideas about science and scientific modeling in physics]. *Revista Electrónica de Investigación en Educación en Ciencias* [*Electronic Journal of Research in Science Education*], *6*(1), 43-60.

Villablanca, S., Montenegro, M., & Ramos-Moore, E. (2020). Analysis of student perceptions of scientific models: Validation of a Spanish-adapted version of the Students' Understanding of Models in Science instrument. *International Journal of Science Education, 42*(17), 2945-2958. https://doi.org/10.1080/09500693.2020.1843735

Wei, S., Liu, X., & Jia, Y. (2014). Using Rasch measurement to validate the instrument of students' understanding of models in science (SUMS). *International Journal of Science and Mathematics Education, 12,* 1067-1082. https://doi.org/10.1007/s10763-013-9459-z

Wright, B. D., & Linacre, J. M. (1994). Reasonable mean-square fit values. *Rasch Measurement Transactions, 8*(3), 370.

APPENDIX A

Conamo Questionnaire

1. Un modelo tiene que ser completamente igual que el objeto que representa, pero a un tamaño diferente [A model must be the same as the object it represents, but on a different scale].
2. Cada modelo tiene que ser idéntico al objeto o fenómeno que estudia [Every model must be identical to the object or phenomenon it studies].
3. Un modelo tiene que ser igual que el objeto que estudia [A model must be the same as the object it studies].
4. Un modelo puede ser un poco diferente del objeto estudiado [A model can differ a little from the object studied].
5. Un modelo y el fenómeno que representa pueden ser diferentes en algunas cosas [A model and the phenomenon it represents can differ in some aspects].
6. Los modelos sirven para poner a prueba las teorías [Models serve to put theories to the test].
7. Los modelos se usan para ayudar a elaborar ideas y teorías sobre los fenómenos científicos [Models are used to help develop ideas and theories about scientific phenomena].
8. Un modelo sirve de poco para plantear preguntas investigables [A model is of little use for posing research questions].
9. Un modelo sirve de poco para elaborar teorías [A model is of little use for developing theories].
10. Los modelos sirven para comprender los fenómenos naturales [Models help us to understand natural phenomena].
11. Si tenemos dos modelos del mismo fenómeno, solo uno resultará útil para entenderlo [If we have two models of the same phenomenon, only one will be useful for understanding it].
12. Los científicos suelen recurrir a varios modelos para estudiar diferentes aspectos de un mismo fenómeno [Scientists usually rely on several models to study different aspects of the same phenomenon].
13. Un solo modelo es siempre suficiente para comprender un fenómeno [A single model is always sufficient for understanding a phenomenon].
14. Se necesitan varios modelos para mostrar distintos puntos de vista de un fenómeno [Several models are needed to show different perspectives on a phenomenon].
15. Para cada fenómeno solo se debe emplear un modelo [Only one model should be used for each phenomenon].
16. Un modelo nunca es definitivo [A model is never definitive].
17. Cuando un modelo es aceptado por los científicos ya nunca es cambiado por otro modelo [When a model is accepted by scientists it is never replaced by another model].
18. Si aparecen nuevos datos o nuevas ideas, los modelos pueden cambiar [If new data emerge or new ideas develop, then models may change].
19. Los modelos que aparecen hoy en los libros son ya definitivos y no cambiarán en el futuro [The models that are currently described in textbooks are definitive and will not change in the future].
20. Todo modelo puede cambiar con el tiempo [Any model can change over time].

APPENDIX B

Instructions for Instrument Administrators

1. Explain that CoNaMo questionnaire participation is voluntary, and their answers will be kept confidential.
2. Explain clearly that the purpose of the questionnaire is find out what students think about models and their role in science. The assessment will not influence their school grades in any way.
3. Students should be seated with some space between them to allow for privacy.
4. Distribute the answer sheets.
5. The questionnaire administrators should not clarify words or phrases or define any of the terms used in the survey. If there are questions about any item, simple respond; just answer the question as you interpret it.
6. No time limit must set for completing the questionnaire.
7. Express your gratitude to the respondents when he/she complete the questionnaire.

APPENDIX C

Instructions for Instrument Rating

1. Students' responses to the CoNaMo are coded on a scale from zero (corresponding to strongly disagree) to four (strongly agree).
2. Scores on reverse worded items must be recoded, inverting the values.
3. A blank response can be considered an indecisive response. Therefore, it can be coded with the value two of the Likert scale.
4. A result for each subscale can be obtained by calculating the sum of individual item results, thus obtaining a scale ranging from 0 to 20: beyond exact replicas (items 1 to 5), purpose of models (items 6 to 10), multiple models (items 11 to 15), and changing models (items 16 al 20).
5. These scores can be recoded and expressed on an interval from zero to one, just dividing by 20.
6. A score of 15 points (.75 if the scale is normalized to a range from zero to one), could serve to differentiate between students with an adequate versus an insufficient understanding of the nature of models and their uses in science.

◆❖◆