Research Article

Marvin Rost*, Ines Sonnenschein, Stephanie Möller and Anja Lembens

Don’t we know enough about models? Integrating a replication study into an introductory chemistry course in higher education

https://doi.org/10.1515/cti-2022-0032
Received October 28, 2022; accepted July 12, 2023; published online September 15, 2023

Abstract: This paper presents the German translation and replication of the Students’ Understanding of Models in Science (SUMS) instrument, aiming to assess how first-semester university students comprehend the submicroscopic level in chemistry courses. The assessment of students’ understanding is a prerequisite for improving teaching practices, particularly in addressing the persistently high drop-out rates observed in chemistry and chemistry-related programs. Employing a quantitative methodology, a sample of 181 undergraduate chemistry students was surveyed. The data were analyzed using structural equation modeling, resulting in two statistical models that demonstrated an excellent fit to the data, although no empirical preference could be established for one model over the other. Based on the investigation, framing models as exact replicas of the natural world cannot be considered an empirically meaningful dimension of understanding models in science. Additionally, the reliabilities of the latent constructs were found to be insufficiently low to establish generalizable measurements. These findings are discussed with a focus on epistemology and advocate for a stronger integration of model theory in chemistry teaching and learning. Finally, the importance of establishing a stronger connection between empirical evidence and the implementation of curricular changes in higher education is emphasized.

Keywords: higher education; chemistry; meta-modeling knowledge

1 Introduction

The drop-out rates observed in undergraduate STEM tracks have reached unacceptable levels. In Germany, it has been reported that approximately 30–40 % of STEM students do not successfully complete their studies (Heublein et al., 2022). Reducing these percentages is of utmost importance, particularly from the standpoint of a technology-and-knowledge-based national economy. Establishing and maintaining a supportive infrastructure that facilitates the development of a sufficient number of next-generation engineers in high-technology environments is crucial.

To ensure that university administration and teaching staff can implement appropriate measures to retain their students, it is imperative for them to comprehend the subject-specific causal factors influencing students’ drop-out rates. A specific challenge within the field of chemistry pertains to universities’ ability to mitigate the

***Corresponding author: Marvin Rost**, Austrian Educational Competence Centre Chemistry, University of Vienna, Porzellangasse 4/2/2, Vienna, 1090, Austria, E-mail: marvin.rost@univie.ac.at. https://orcid.org/0000-0002-9580-2035
Ines Sonnenschein, Wandelwerk, University of Applied Sciences Muenster, Johann-Krane-Weg 21, 48149 Muenster, Germany
Stephanie Möller, Department of Chemical Engineering, University of Applied Sciences Muenster, Stegerwaldstraße 39, 48565 Steinfurt, Germany
Anja Lembens, Austrian Educational Competence Centre Chemistry, University of Vienna, Porzellangasse 4/2/2, Vienna, 1090, Austria

drop-out rate resulting from failed exams in introductory chemistry lectures (Fleischer et al., 2019). By offering empirical evidence of these causal factors to teachers at universities and schools, chemistry education research can equip them with the necessary tools to effectively utilize their valuable time and resources in activities that promote meaningful learning. The central focus of this learning process should be the connection between the particle level and observable phenomena, which constitutes the cornerstone of chemistry knowledge (Reid, 2021a; Stowe et al., 2021). In order to facilitate this connection, a solid understanding of scientific modeling is advantageous for both students and teachers (Ke & Schwarz, 2021). Thus, it becomes important to cultivate students' comprehension of scientific models in general, as this will ultimately contribute to the development of a stronger understanding of chemistry (Schwedler & Kaldewey, 2020). Considering this line of reasoning, it becomes apparent that assessing students' understanding of scientific models becomes a necessary prerequisite. This is particularly relevant since the learning environments in chemistry education are not well-suited to complement their curriculum with individual tutoring on models and modeling in science.

The *Students' Understanding of Models in Science* instrument (SUMS, Treagust et al., 2002) is considered a potential candidate for such an assessment due to its ability to strike a balance between meaningfulness and efficiency when evaluating students' comprehension of scientific models. In the subsequent sections, this article presents the German translation of SUMS and describes the validation process.

2 Theoretical background

2.1 Epistemological considerations

The exploration of models and modeling in chemistry is inherently tied to epistemological considerations (Erduran, 2001; Klein, 2003). Simultaneously, within science education, there exists a tendency to view models as both direct representations of target systems and as subject-oriented processes for making sense of these target systems (Rost & Knuuttila, 2022). However, this conflation of perspectives is contradictory. On one hand, when a scientist attempts to depict an atom and argues for a structural relationship within the representational vehicle (e.g., a sketched circle with dots), there is no direct means of justifying that specific structure. On the other hand, if the scientist posits that a model retains value as long as it accurately represents the phenomenon under investigation (Stowe & Esselman, 2022), one could question how this approach generates subject-specific knowledge in the first place.

This article's position aligns with the notion of models as epistemic artifacts (Knuuttila, 2011, 2021). This perspective alleviates the burden of striving for more or less precise representational endeavors or falling into the trap of regarding models solely as practical constructions. Instead, it allows for the construction of scientific knowledge.

2.2 Empirical considerations

In order to effectively mitigate drop-outs in chemistry courses, a thorough understanding of students' comprehension of models and modeling, as well as its connection to their prior knowledge, is imperative (Chittleborough & Treagust, 2007). Merely assessing students' understanding of models is far from adequate when it comes to teaching any science subject (Schwarz et al., 2022); however, it is a mandatory requirement (Constantinou et al., 2019). This necessity arises from two primary reasons. Firstly, Fleischer et al. (2019) demonstrate a direct influence of prior knowledge on students' intentions to discontinue their studies. Secondly, the comprehension of the particle level precedes the acquisition of chemistry knowledge (Sumfleth & Nakoinz, 2019). Regrettably, the endeavor to integrate prior subject knowledge, and knowledge of models and modeling within the teaching process encounters various challenges, encompassing the cognitive demands placed on learners (Johnstone, 1993) and the epistemological perspectives of teachers (Oh & Oh, 2011). These challenges accumulate, contributing to the

inherent difficulty of learning chemistry (Reid, 2021b). In conclusion, without gaining insights into how to engage with students' preconceptions (Duit & Treagust, 2003), lecturers risk merely reproducing a superficial presentation of the philosophy of science during introductory chemistry lectures.

2.3 The SUMS inventory

In order to systematically comprehend students' understanding of models, the utilization of an assessment tool becomes necessary for both researchers and teachers. Particularly in the context of university settings with large and diverse first-year cohorts, where individual assessment and acknowledgment of students' perspectives are resource-intensive, a quantitative instrument would prove beneficial. The starting point for this endeavor is the SUMS inventory (Treagust et al., 2002), which has gained recognition as a questionnaire designed to quantitatively assess students' comprehension of the properties and applications of models in science (Mathesius & Krell, 2019). Distinct from performance-oriented approaches such as the Framework for Modeling Competence (Constantinou et al., 2019; Upmeier zu Belzen et al., 2019), SUMS captures learners' perspectives regarding models and their utilization by scientists. In this regard, Treagust et al. (2002) propose five latent factors to elucidate the understanding of models.

(1) *Models as multiple representations* (MR) refers to the diverse forms that models can assume, illustrating how distinct model objects can convey a single target system. MR endeavors to encompass the range of representational variations associated with a particular target system. Example: "Many models represent different versions of the phenomenon."
(2) *Models as exact replica* (ER) refers to assessing the perceived proximity between a model object and its corresponding target system. ER entails establishing a structural connection between the model's structure and that of the target system, thus determining the degree of structural resemblance. Example: "A model shows what the real thing does and what it looks like."
(3) *Models as explanatory tools* (ET) refers to the capacity of model objects to facilitate understanding of a target system by enabling access to non-visible objects and processes. ET serves to articulate learners' comprehension through the utilization of a model as a representational means. Example: "Models help create a picture in your mind of the scientific happening."
(4) *Uses of scientific models* (USM) refers to the processes employed by scientists when utilizing models, specifically focusing on how models and modeling are implemented in scientific practice. USM encompasses the diverse ways in which models are employed within the scientific practices of various fields. Example: "Models are used to make and test predictions about a scientific event."
(5) *Changing nature of models* (CNM) refers to the capacity of model objects to be revised, indicating that models are not deterministic but rather susceptible to modification in response to new data or invalidated predictions. CNM encompasses information regarding the degree to which learners perceive models as either permanent or subject to change. Example: "A model can change if there are new findings."

The SUMS inventory has recently garnered attention through replication studies conducted in the USA and Chile, resulting in mixed findings. Villablanca et al. (2020) present a successful replication of the original structure after translating the questionnaire into Spanish. They suggest further research to better comprehend the variance within the respective factors. However, their use of Cronbach's α as a reliability measure has faced substantial criticism in the empirical literature over the past decade (Hayes & Coutts, 2020; Padilla, 2019; Taber, 2018; Yang & Green, 2011).

In contrast to the successful replication in Spanish, Lazenby and Becker (2021) conclude that there is insufficient evidence to support the use of the SUMS inventory in its original form as an assessment tool. They were unable to replicate the original factor structure and reported cross-loadings during their exploratory factor analysis. Despite making theoretically justified modifications to the scales, they were still required to reject their

measurement models. These findings clearly indicate the limited transferability of the instrument, leading the authors to caution researchers and practitioners against using the SUMS instrument in contexts outside of its original development without additional evidence of validity and reliability.

Therefore, the main objective of this article is to fulfill this need by providing an accurate account of the translation process of the SUMS inventory, along with modifications that maintain sufficient fidelity to the original instrument. This approach ensures a transparent proposal and analysis of the psychometric structure. The research questions and proposed structure are presented as follows:

2.4 Research questions

(1) Can the factor structure of the SUMS inventory, as originally proposed by Treagust et al. (2002), be replicated in a translated (German) version?
(2) If the proposed factor structure holds, to what extent do the respective scales demonstrate the necessary reliability to serve as valid measurements for subsequent analyses?

2.5 Hypothetical psychometric structure

Four distinct models were agreed upon to integrate the aforementioned information into a testable empirical framework (Figures 1–4). The first model replicates the structure proposed by Treagust et al. (2002), as the inconclusive findings from prior research resulted in retaining the assumption of five correlated factors. The second model was derived from theoretical considerations regarding the ER factor. The empirical challenges encountered thus far in relation to ER are theoretically justified, given the intention to conceptualize models beyond mere depictions. For example, "Many models are used to show how it depends on individual's different ideas on what things look like or how they work." (MR) should not correlate with "A model shows what the real thing does and what it looks like." (ER). Hence, a proposal is put forth to test a structure comprising of only four correlated factors, namely MR, ET, USM, and CNM. The remaining two models represent expanded versions of the aforementioned structures, encompassing their respective five or four factors in addition to an overarching SUMS factor. By incorporating a higher-order factor model (Brunner et al., 2012), this approach signifies the assumption of a comprehensive factor of general model understanding. This overarching factor comprises subfactors that align with similar yet discernible directions, thereby reflecting Treagust et al. (2002) foundational assertion regarding students' comprehension of models: "[They] have their own personal and unique understanding of the role of scientific models in science built up through their life experiences." (p. 358)

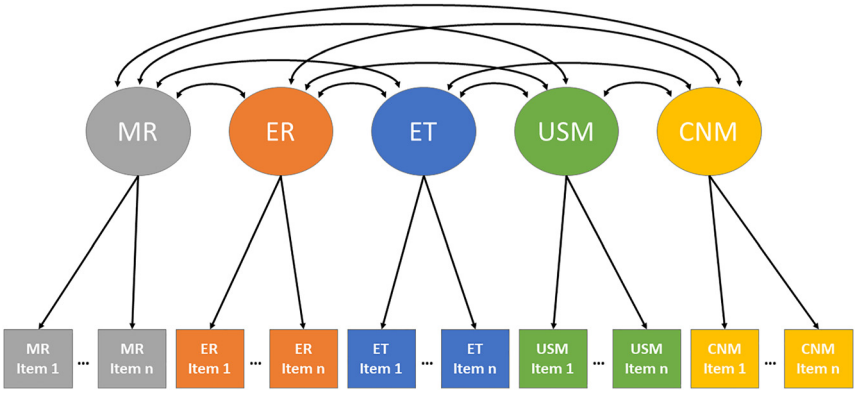

Figure 1: Reproduced psychometric assumption of SUMS by Treagust et al. (2002). Five correlated factors constitute views on models and modeling.

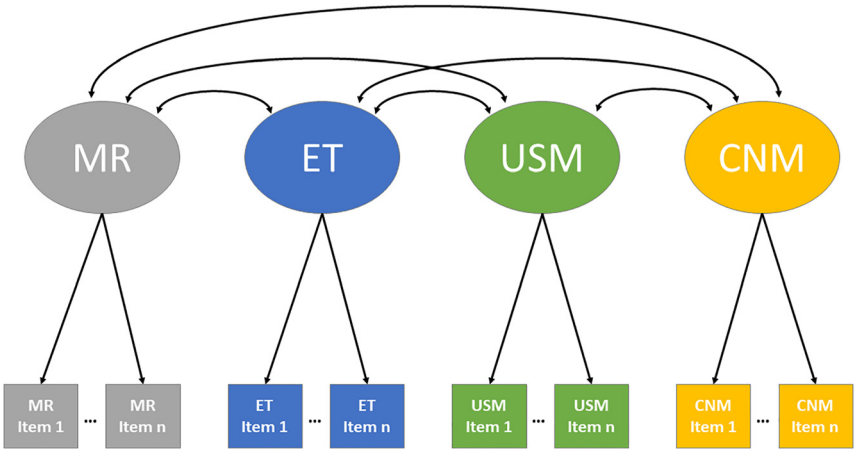

Figure 2: Adapted psychometric assumption of SUMS by Treagust et al. (2002). Excluding ER leads to four correlated factors for views on models and modeling.

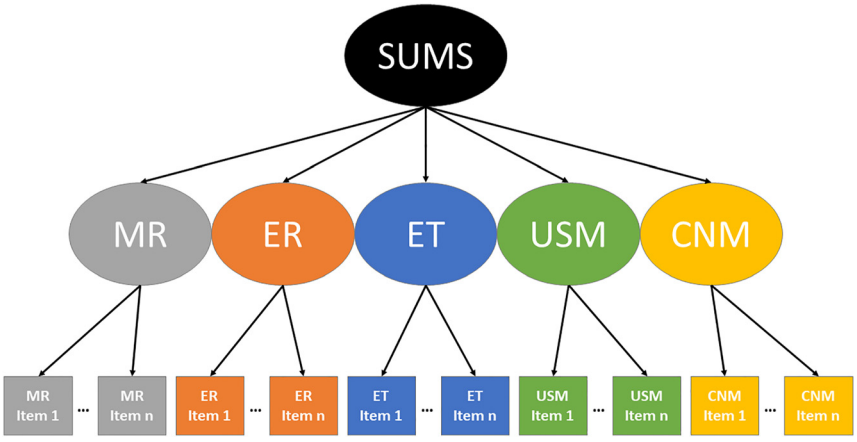

Figure 3: Extended psychometric assumption of SUMS following Treagust et al. (2002). Five non-correlated factors are connected via an overarching general factor.

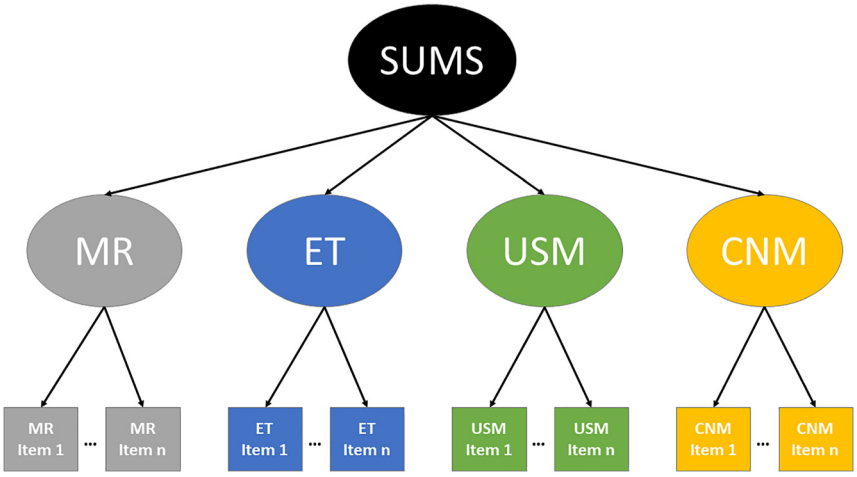

Figure 4: Extended and adapted psychometric assumption of SUMS following Treagust et al. (2002). Excluding ER and introducing an overarching general factor constitute the theoretical assumption of views on models and modeling.

3 Design and methods

3.1 Context

This investigation contributes to a German university's empirically grounded program aimed at facilitating meaningful comprehension of the submicroscopic level among first-semester students. The primary objective is to minimize the influence of prior knowledge on academic success in chemistry (Fleischer et al., 2019; Hailikari & Nevgi, 2010). Given the constraints of limited resources for individualized teaching in large cohorts, it becomes essential to establish efficient and generalizable measures that can be integrated into the learning environments.

With respect to chemistry-related lectures, the university has made the decision to review the content of their lectures based on empirical research in chemistry education. Consequently, the teaching staff aims to enhance their understanding of how students grasp the particle level in the context of scientific modeling, moving beyond the mere presentation of declarative knowledge and fostering a proactive learning experience.

3.2 Translation process

For the translation process, three individuals were involved, two of whom translated the items into German. One of these individuals was the first author of this paper, while the other was a native English speaker. Additionally, a German native speaker with an academic degree in English contributed to the translation process. Following this, the items were re-translated into German by the native English speaker, and deviations were carefully examined. Multiple rounds of discussion were conducted, which did not yield significant concerns regarding content validity or potential sources of misunderstanding. Throughout the translation process, precautions were taken to address common sources of validity constraints in quantitative tools. For instance, it was ensured that item texts containing conjunctions such as *and* were avoided, in accordance with recommendations by Haladyna et al. (2002). As a result, the item pool became more extensive than the original instrument, as items containing conjunctions were split into two separate items.

3.3 Target group & data collection

Following the transition of the questionnaire into an online format, data were collected from a total of 181 first-year students enrolled in three distinct study programs: physics (N = 60), dietetics (N = 79), and chemistry engineering (N = 42). The digital questionnaire was completed by the students using their personal devices during a lecture session, in the presence of the lecturer.

3.4 Data analysis

After extracting the data from the university's database, the obtained Excel sheet was transferred to the statistical software R (R Core Team, 2020) for further analysis. An exploratory data analysis was initiated (Bryer & Speerschneider, 2016; Lüdecke et al., 2022; Revelle, 2019) with a specific focus on examining item correlations within the five hypothesized factors. The reporting of these correlations adheres to the guidelines outlined by Funder and Ozer (2019). Subsequently, confirmatory factor models were estimated and evaluated, which were then summarized in structural equation models (Hu & Bentler, 1999; Rosseel, 2012; Ziegler & Hagemann, 2015).

4 Results

Following the exclusion of items with non-significant correlations, the remaining items were deemed to measure the latent constructs (SUMS factors) with different reliabilities and an additive constant for each item (Bühner, 2021). Consequently, the four dimensions, namely MR, ET, CNM, and USM, exhibited a satisfactory fit when considering essentially tau equivalent measurement models, as presented in Table 1. After exploring various measurement models, ER was ultimately excluded from further analysis. Neither could an appropriate measurement model be identified, nor did the items exhibit correlations that would permit the establishment of a shared scale. Figure 5 displays the distribution of answers for the ER factor. It is important to highlight the varying directions of agreement observed for certain items, suggesting a lack of fit at the descriptive level. For example, the product-moment correlation between ER_9 ("A model should be an exact replica.") and ER_12 ("Every part of a model should clearly show what it represents.")[1] is negative, statistically not significant, and tiny ($r = -0.04$, 95 % CI [−0.18, 0.11], $t(179) = -0.53$, $p < 0.600$).

The descriptive answer distributions of the remaining four factors are presented in the Appendix section of this article. After the exclusion of ER, the resultant structural equation model with a general factor demonstrated a good fit ($\chi^2 = 167.9$ (162), $p = 0.36$; CFI = 0.98; RMSEA = 0.01, 95 % CI [0; 0.04]; SRMR = 0.07). The structure is depicted in Figure 6. One item pair was allowed to share error variance: "Models show different perspectives on scientific phenomena" (MR5) and "Models are changed if new theories disagree with them" (CNM25). Both items are related to the conditional value of models, indicating that the introduction of a new theory yields fresh insights into the subject under consideration. Furthermore, the hypothesized model with four correlated factors and the absence of a general factor also demonstrated a good fit ($\chi^2 = 169.3$ (161), $p = 0.31$; CFI = 0.97; RMSEA = 0.02, 95 % CI [0; 0.04]; SRMR = 0.07).

The range of Cronbach's α values for our items was found to be acceptably high ($\propto_{\text{Cronbach}} \approx 0.8$). Nevertheless, upon employing state-of-the-art criteria recommended for assessing the reliability of the instrument (Hayes & Coutts, 2020; Padilla, 2019), less reliable estimates for the latent constructs were uncovered (Table 2).

5 Discussion

Based on the observed misfit between the empirical data and the hypothesized ER factor, it can be concluded that the translated version of the SUMS inventory does not accurately reflect the original publication (Treagust et al., 2002).

No empirically conclusive reasons were found to differentiate between the psychometric structure comprising the four separate factors (MR, ET, CNM, USM) that demonstrated a good fit and the structure incorporating a general factor representing an overall estimate of students' model understanding.

Table 1: Fit measures for essentially tau equivalent measurement models of the five dimensions. Dimension ER does not show sufficient fit measures.

Fit measure	MR	ER	ET	CNM	USM
Degrees of freedom (df)	9	9	8	5	9
Chi-square (χ^2)	10.11	18.4	8.68	3.71	6.14
p-value	0.34	<0.05	0.47	0.59	0.73
Comparative Fit Index (CFI)	0.97	0.82	0.99	0.99	0.99
RMSEA	0.03	0.08	0	0	0
SRMR	0.07	0.07	0.05	0.05	0.05

1 Please note that the item phrasings presented throughout the article's empirical part are non-validated re-translations of the German item versions by the first author for illustrative purposes only.

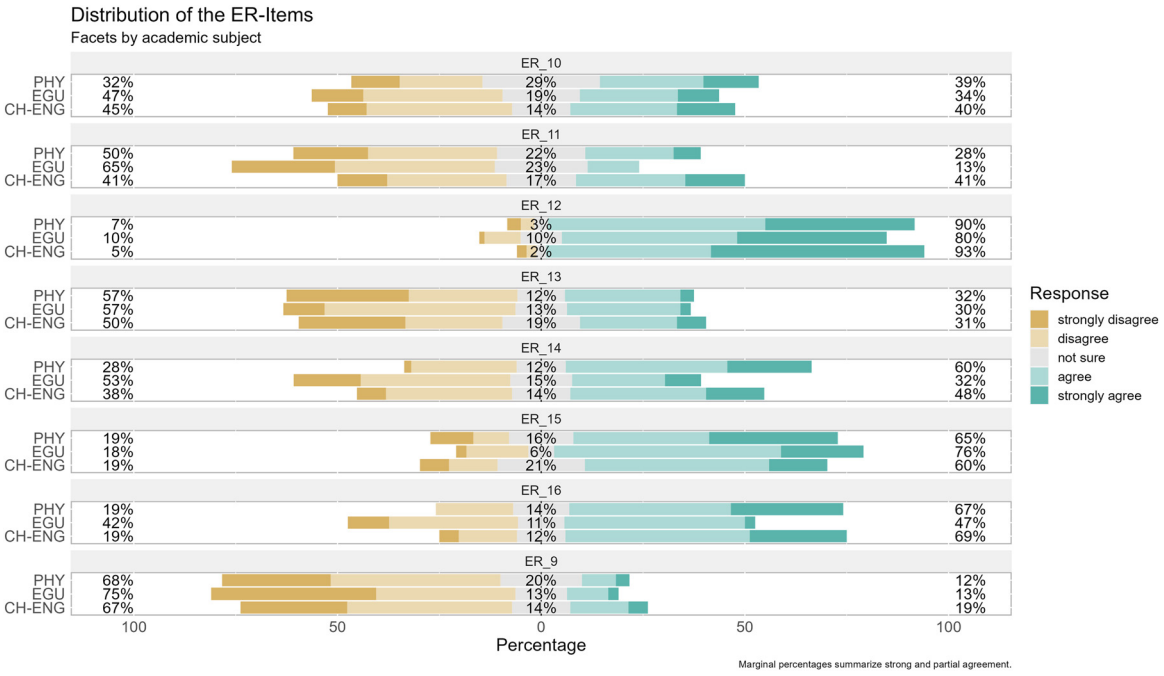

Figure 5: Distribution of participants' answers to the ER items. The figure differentiates between the three academic tracks (PHY, physics; EGU, dietetics; CH-ENG, chemistry engineering).

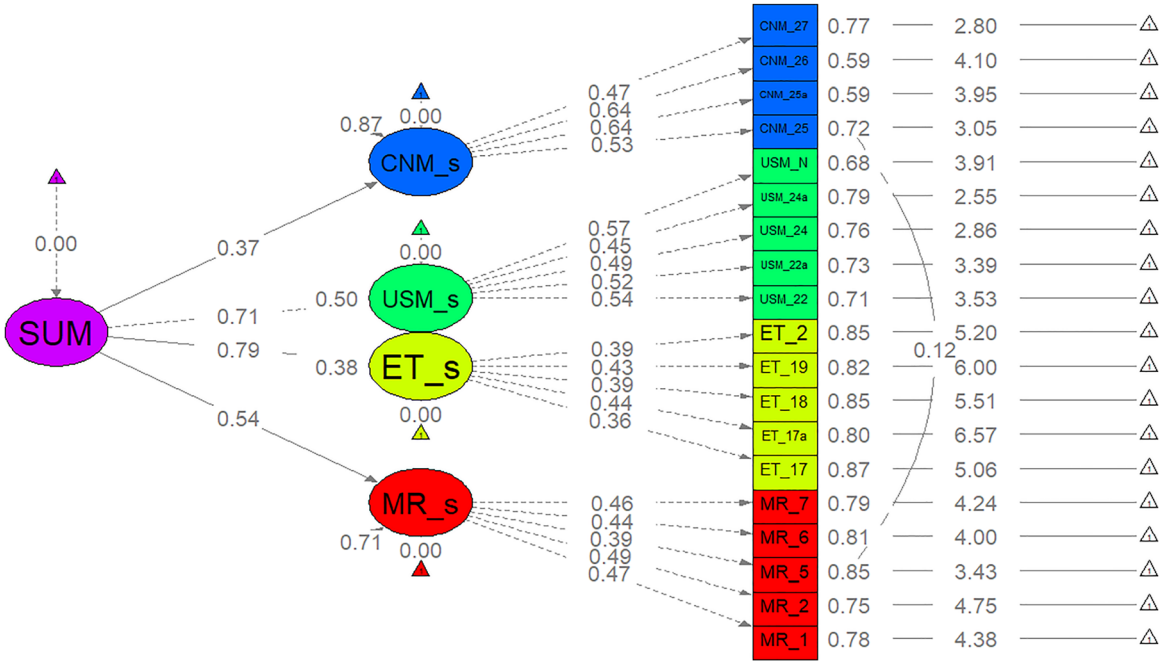

Figure 6: Four-factor structure, excluding ER and including a general factor. $\chi^2 = 168.0$, $p = 0.36$; CFI = 0.98, RMSEA = 0.01, 95 % CI [0; 0.04]; SRMR = 0.07.

An appropriate reliability range could not be established, and concerns arise regarding the item correlations within the nomological net (Ziegler, 2014; Ziegler & Hagemann, 2015). Treagust et al. (2002) already highlighted the problematic nature of understanding models as exact replicas. Both this investigation and the study by Lazenby

Table 2: McDonald's omega (Hayes & Coutts, 2020) with 95 % confidence intervals.

Factor	ω	CI (low)	CI (high)
MR	0.593	0.462	0.675
ET	0.617	0.495	0.697
USM	0.665	0.564	0.732
CNM	0.669	0.561	0.751

and Becker (2021) were unable to establish a measurement model for this particular dimension, despite the original study and the Spanish replication (Villablanca et al., 2020) reporting a fit between the data and the theory. However, each of these studies encountered challenges regarding the validity of the significant correlations among the five factors.

If scientific models are considered as enhancing understanding of target systems through multiple theoretical expressions, it follows that conceptualizing models as mere copies of reality should not exhibit correlation. However, undergraduate students consistently express agreement with items from both the ER and MR dimensions in the same direction. It appears that these students possess an eclectic understanding of models, potentially stemming from the conflation of mutually exclusive theoretical perspectives prevalent in science education research and teaching (Rost & Knuuttila, 2022). In this field, models are often treated simultaneously as structural representations and as tools, contingent upon subjective judgments by users regarding their purpose, context, and timing. At this stage, it is not appropriate to endorse one perspective over the other. However, it should be acknowledged that such conflation is deemed inconsistent (Knuuttila, 2011) and has the potential to result in misunderstandings among stakeholders in science education.

6 Conclusions

On one hand, a valid instrument to effectively capture students' understanding of models with appropriate reliability could not be established. Consequently, drawing generalizable empirical conclusions regarding the connection between model understanding and knowledge gains in general chemistry lectures remains challenging. On the other hand, the development of such an instrument remains an urgent matter. Understanding learners' conceptions of how scientists acquire knowledge is a crucial aspect of future-oriented science education (Oh & Oh, 2011; Schwarz & White, 2005; Schwarz et al., 2022). In empirical investigations, it becomes necessary to devise approaches that integrate qualitative data, establish case studies, and foster a qualitative understanding of students' modeling practices (Göhner et al., 2022; Schwarz et al., 2009). While investigating verbal descriptions or drawings in themselves is valuable, these learning artifacts also serve as promising candidates for integration with quantitative approaches. This methodological diversity has gained traction, including the application of tools from the fields of machine learning and natural language processing (Rost, 2022; Zhai et al., 2022). This article contributes a quantitative perspective to expand beyond the assessment of students' performance (Schwarz et al., 2022), thereby enhancing the understanding of how students comprehend models and modeling in science education.

The findings underscore the significance of a nuanced understanding of the epistemological perspective on models and modeling. Students exhibit a conflation of the multiplicity of models with the notion of models as exact replicas across all investigated studies, demonstrating a consistent inconsistency. Therefore, it is imperative for chemistry educators to provide students with a strong knowledge foundation and an awareness of how scientists construct knowledge since knowledge, without an understanding of its origins, loses its status as such (Hofer, 2002; Sendur et al., 2017).

Lastly, the scientific community is invited to share their concerns and suggestions for refinement, embrace empirically driven approaches in the field, and contribute to the development of a generalizable assessment tool for evaluating students' understanding of models in chemistry-related learning and teaching.

Appendix

The following Figures 7–10 contain the distribution of answers to all items. The original SUMS assignments cluster them into the respectively presented factors.

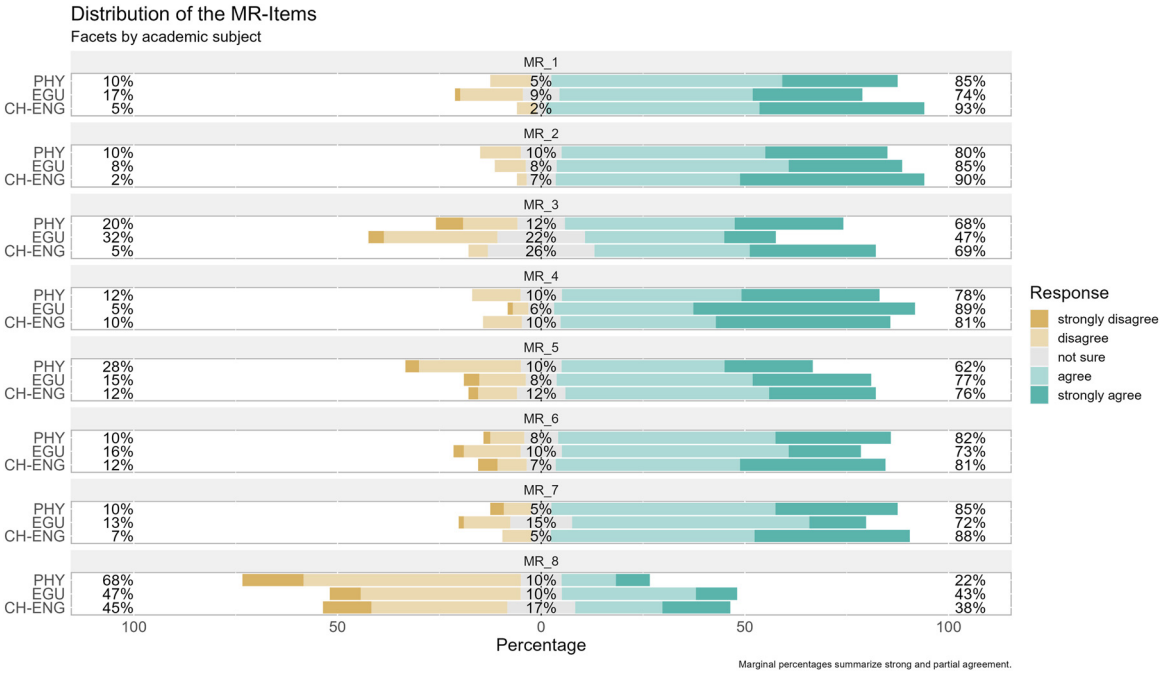

Figure 7: Distribution of participants' answers to the MR items. The figure differentiates between the three academic tracks (PHY, physics; EGU, dietetics; CH-ENG, chemistry engineering).

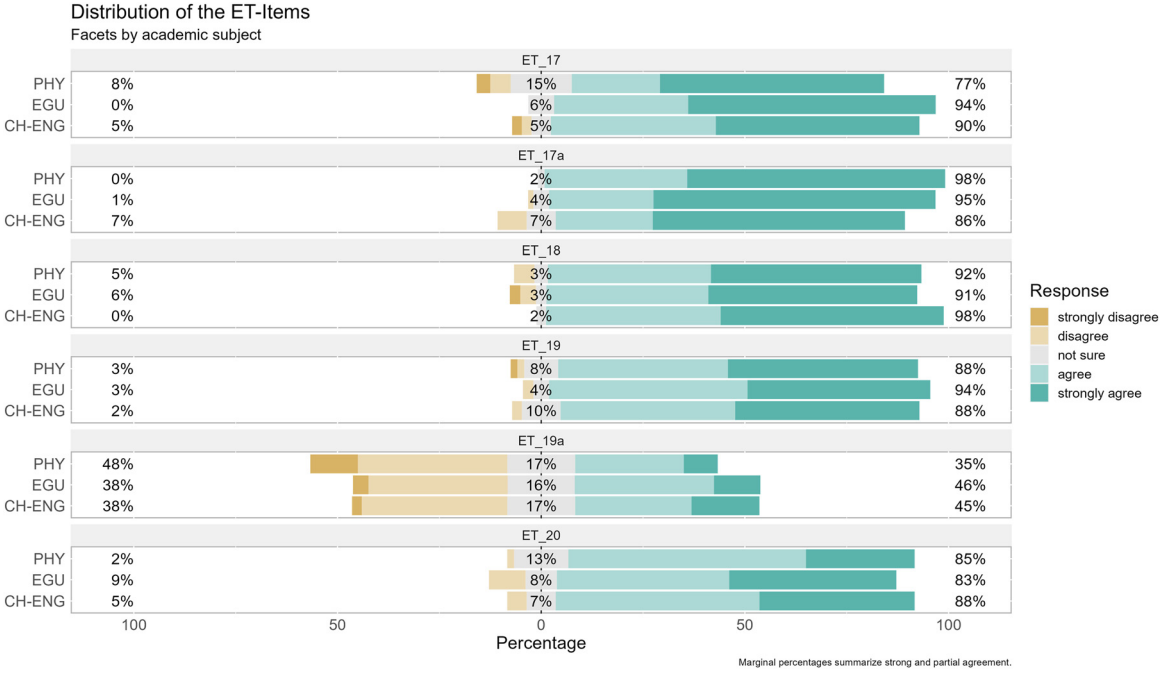

Figure 8: Distribution of participants' answers to the ET items. The figure differentiates between the three academic tracks (PHY, physics; EGU, dietetics; CH-ENG, chemistry engineering).

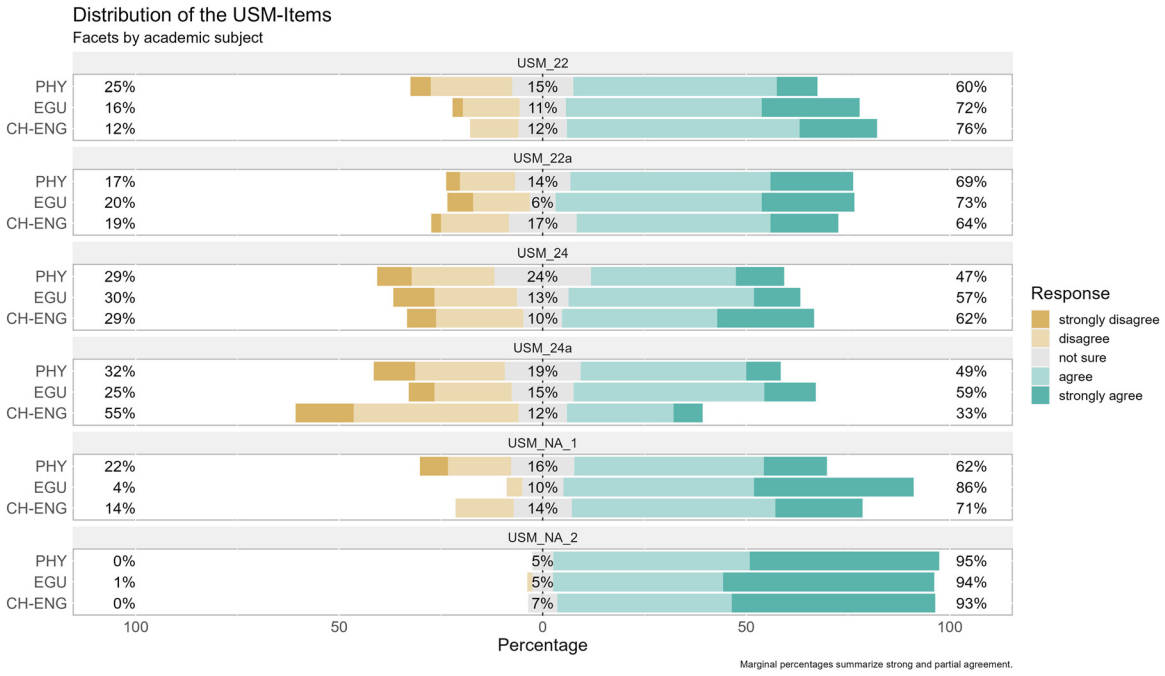

Figure 9: Distribution of participants' answers to the USM items. The figure differentiates between the three academic tracks (PHY, physics; EGU, dietetics; CH-ENG, chemistry engineering).

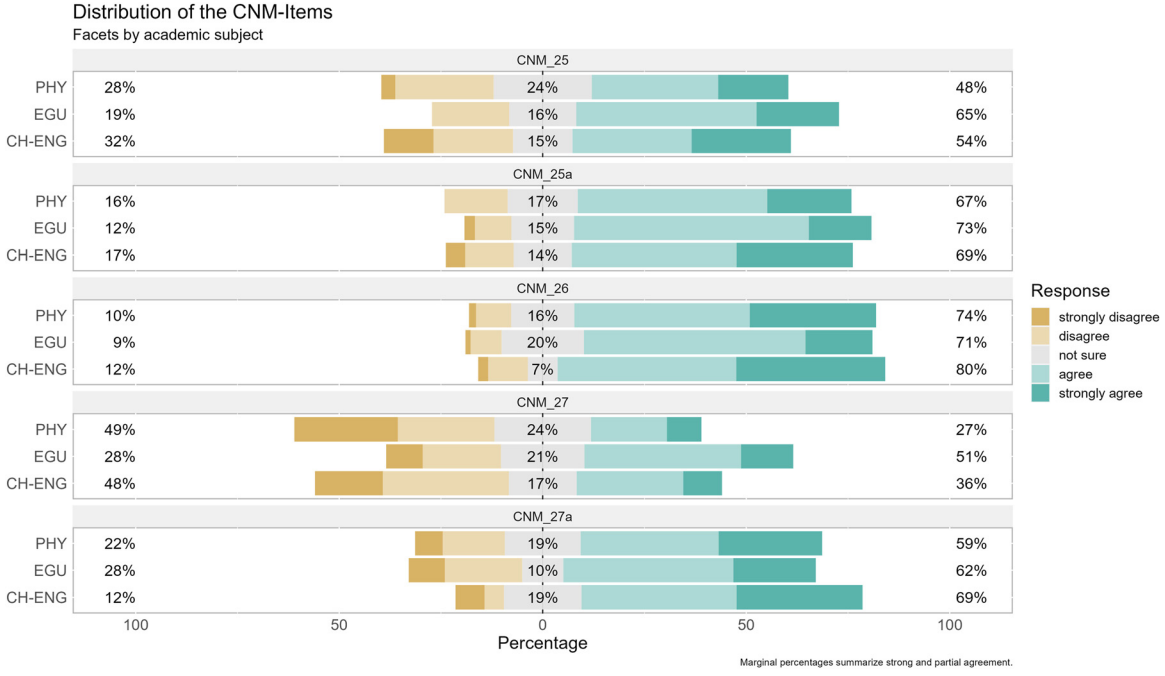

Figure 10: Distribution of participants' answers to the CNM items. The figure differentiates between the three academic tracks (PHY, physics; EGU, dietetics; CH-ENG, chemistry engineering).

Research ethics: Antecedent to the study, explicit participant consent was secured. Subsequently, data was anonymized with individual codes for potential erasure. During analysis, these codes were replaced by sequential numbers by the data collector, ensuring analyst access without compromising raw data security.

Author contributions: Conceptualization: MR, IS, SM, AL; Methodology: MR, IS; Investigation: IS, SM; Data Curation: IS, MR; Formal Analysis: MR; Writing – Original Draft: MR; Writing – Review & Editing: MR, IS, SM, AL. The authors have accepted responsibility for the entire content of this manuscript and approved its submission.
Competing interests: The authors state no conflict of interest.
Research funding: None declared.
Data availability: The data security statement precludes sharing the raw data.

References

Brunner, M., Nagy, G., & Wilhelm, O. (2012). A tutorial on hierarchically structured constructs: Hierarchically structured constructs. *Journal of Personality*, *80*(4), 796–846.

Bryer, J., & Speerschneider, K. (2016). likert: Analysis and visualization of likert based items. https://CRAN.R-project.org/package=likert

Bühner, M. (2021). *Einführung in die Test- und Fragebogenkonstruktion [Introduction to test and questionnaire design]*. ps Psychologie (4., korrigierte und erweiterte Auflage.). Pearson.

Chittleborough, G., & Treagust, D. F. (2007). The modelling ability of non-major chemistry students and their understanding of the sub-microscopic level. *Chemical Education Research and Practice*, *8*(3), 274–292.

Constantinou, C. P., Nicolaou, C. T., & Papaevripidou, M. (2019). A framework for modeling-based learning, teaching, and assessment. In A. Upmeier zu Belzen, D. Krüger & J. van Driel (Eds.), *Towards a competence-based view on models and modeling in science education* (pp. 39–58). Springer International Publishing.

Duit, R., & Treagust, D. F. (2003). Conceptual change: A powerful framework for improving science teaching and learning. *International Journal of Science Education*, *25*(6), 671–688.

Erduran, S. (2001). Philosophy of chemistry: An emerging field with implications for chemistry education. *Science and Education*, *10*(6), 581–593.

Fleischer, J., Leutner, D., Brand, M., Fischer, H., Lang, M., Schmiemann, P., & Sumfleth, E. (2019). Vorhersage des Studienabbruchs in naturwissenschaftlich-technischen Studiengängen [Predicting dropout in science and engineering courses in higher education]. *Zeitschrift für Erziehungswissenschaft*, *22*(5), 1077–1097.

Funder, D. C., & Ozer, D. J. (2019). Evaluating effect size in psychological research: Sense and nonsense. *Advances in Methods and Practices in Psychological Science*, *2*(2), 156–168.

Göhner, M. F., Bielik, T., & Krell, M. (2022). Investigating the dimensions of modeling competence among preservice science teachers: Meta-modeling knowledge, modeling practice, and modeling product. *Journal of Research in Science Teaching*. https://onlinelibrary.wiley.com/doi/10.1002/tea.21759

Hailikari, T. K., & Nevgi, A. (2010). How to diagnose at-risk students in chemistry: The case of prior knowledge assessment. *International Journal of Science Education*, *32*(15), 2079–2095.

Haladyna, T. M., Downing, S. M., & Rodriguez, M. C. (2002). A review of multiple-choice item-writing guidelines for classroom assessment. *Applied Measurement in Education*, *15*(3), 309–333.

Hayes, A. F., & Coutts, J. J. (2020). Use omega rather than Cronbach's alpha for estimating reliability. But…. *Communication Methods and Measures*, *14*(1), 1–24.

Heublein, U., Hutzsch, C., & Schmelzer, R. (2022). Die Entwicklung der Studienabbruchquoten in Deutschland [The development of drop-out rates in higher education in Germany]. *DZHW Brief. Deutsches Zentrum für Hochschul- und Wissenschaftsforschung (DZHW)*. https://www.dzhw.eu/publikationen/pub_show?pub_id=7922&pub_type=kbr

Hofer, B. K. (2002). Personal Epistemology as a Psychological and Educational Construct: An Introduction. In *Personal epistemology: The psychology about knowledge and knowing*. Routledge.

Hu, L., & Bentler, P. M. (1999). Cutoff criteria for fit indexes in covariance structure analysis: Conventional criteria versus new alternatives. *Structural Equation Modeling: A Multidisciplinary Journal*, *6*(1), 1–55.

Johnstone, A. H. (1993). The development of chemistry teaching: A changing response to changing demand. *Journal of Chemical Education*, *70*(9), 701.

Ke, L., & Schwarz, C. V. (2021). Supporting students' meaningful engagement in scientific modeling through epistemological messages: A case study of contrasting teaching approaches. *Journal of Research in Science Teaching*, *58*(3), 335–365.

Klein, U. (2003). *Experiments, models, paper tools*. Writing science. Stanford University Press.

Knuuttila, T. (2011). Modelling and representing: An artefactual approach to model-based representation. *Studies in History and Philosophy of Science*, *42*, 262–271.

Knuuttila, T. (2021). Epistemic artifacts and the modal dimension of modeling. *European Journal for Philosophy of Science*, *11*(3), 1–18.

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*(1), 62–76.

Lüdecke, D., Ben-Shachar, M. S., Patil, I., Wiernick, B. M., Bacher, E., Thériault, R., & Makowski, D. (2022). *easystats: Framework for easy statistical modeling, visualization, and reporting*. CRAN. https://easystats.github.io/easystats/

Mathesius, S., & Krell, M. (2019). Assessing modeling competence with questionnaires. In A. Upmeier zu Belzen, D. Krüger & J. van Driel (Eds.), *Towards a Competence-Based View on Models and Modeling in Science Education*, Models and Modeling in Science Education (Vol. 12, pp. 117–131). Springer International Publishing.

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.

Padilla, M. A. (2019). A primer on reliability via coefficient alpha and omega. *Archives of Psychology*, *3*(8), 1–15.

R Core Team (2020). *R: A language and environment for statistical computing*. R Foundation for Statistical Computing. https://www.R-project.org

Reid, N. (2021a). *The Johnstone triangle: The key to understanding chemistry*. Advances in chemistry education series. Royal Society of Chemistry.

Reid, N. (2021b). Johnstone's triangle: Why chemistry is difficult. In N. Reid (Ed.), *The Johnstone triangle: The key to understanding chemistry* (pp. 48–71). Royal Society of Chemistry.

Revelle, W. (2019). *psych: Procedures for psychological, psychometric, and personality research*. https://CRAN.R-project.org/package=psych

Rosseel, Y. (2012). lavaan: An R package for structural equation modeling. *Journal of Statistical Software*, *48*(2), 1–36.

Rost, M. (2022). Do models depict particles? Analyzing Austrian chemistry school textbooks via natural language processing. *CHEMKON*, *29*(S1), 325–330.

Rost, M., & Knuuttila, T. (2022). Models as epistemic artifacts for scientific reasoning in science education research. *Education Sciences*, *12*(4), 276.

Schwarz, C. V., Ke, L., Salgado, M., & Manz, E. (2022). Beyond assessing knowledge about models and modeling: Moving toward expansive, meaningful, and equitable modeling practice. *Journal of Research in Science Teaching*, *59*(6), 1086–1096.

Schwarz, C. V., Reiser, B. J., Davis, E. A., Kenyon, L., Achér, A., Fortus, D., Shwartz, Y., Hug, B., & Krajcik, J. (2009). Developing a learning progression for scientific modeling: Making scientific modeling accessible and meaningful for learners. *Journal of Research in Science Teaching*, *46*(6), 632–654.

Schwarz, C. V., & White, B. Y. (2005). Metamodeling knowledge: Developing students' understanding of scientific modeling. *Cognition and Instruction*, *23*(2), 165–205.

Schwedler, S., & Kaldewey, M. (2020). Linking the submicroscopic and symbolic level in physical chemistry: How voluntary simulation-based learning activities foster first-year university students' conceptual understanding. *Chemistry Education: Research and Practice*, *21*(4), 1132–1147.

Sendur, G., Polat, M., & Kazancı, C. (2017). Does a course on the history and philosophy of chemistry have any effect on prospective chemistry teachers' perceptions? The case of chemistry and the chemist. *Chemistry Education: Research and Practice*, *18*(4), 601–629.

Stowe, R. L., & Esselman, B. J. (2022). The picture is not the point: Toward using representations as models for making sense of phenomena. *Journal of Chemical Education*, *100*(1), 15–21.

Stowe, R. L., Scharlott, L. J., Ralph, V. R., Becker, N. M., & Cooper, M. M. (2021). You are what you assess: The case for emphasizing chemistry on chemistry assessments. *Journal of Chemical Education*, *98*(8), 2490–2495.

Sumfleth, E., & Nakoinz, S. (2019). Chemie verstehen—Beobachtbare makroskopische Phänomene auf submikroskopischer Ebene modellbasiert interpretieren [Understanding chemistry-interpreting observable macroscopic phenomena by reference to the submicroscopic level]. *Zeitschrift für Didaktik der Naturwissenschaften*, *25*(1), 231–243.

Taber, K. S. (2018). The use of Cronbach's alpha when developing and reporting research instruments in science education. *Research in Science Education*, *48*(6), 1273–1296.

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.

Upmeier zu Belzen, A., van Driel, J., & Krüger, D. (2019). Introducing a framework for modeling competence. In A. Upmeier zu Belzen, D. Krüger & J. van Driel (Eds.), *Towards a competence-based view on models and modeling in science education*, Models and modeling in science education (Vol. 12, pp. 3–19). Springer International Publishing.

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.

Yang, Y., & Green, S. B. (2011). Coefficient alpha: A reliability coefficient for the 21st century? *Journal of Psychoeducational Assessment*, *29*(4), 377–392.

Zhai, X., He, P., & Krajcik, J. (2022). Applying machine learning to automatically assess scientific models. *Journal of Research in Science Teaching*, *59*(10), 1765–1794.

Ziegler, M. (2014). Stop and state your intentions: Let's not forget the ABC of test construction. *European Journal of Psychological Assessment*, *30*(4), 239–242.

Ziegler, M., & Hagemann, D. (2015). Testing the unidimensionality of items: Pitfalls and loopholes. *European Journal of Psychological Assessment*, *31*(4), 231–237.