MODESTUM EURASIA Journal of Mathematics, Science and Technology Education, 2026, 22(9), em2911
ISSN:1305-8223 (online)
OPEN ACCESS Research Paper https://doi.org/10.29333/ejmste/19240

General model understanding in physics education: Validating a two-dimensional structure of fidelity of gestalt and functional fidelity

Henry Witteborn [1], Joaquin Veith [1], Malte S. Ubben [1*]

[1] Institut für Didaktik der Physik, Fakultät für Physik und Erdsystemwissenschaften, Leipzig University, Leipzig, GERMANY

Received 27 April 2026 ▪ Accepted 24 July 2026

Abstract
Models are essential to knowledge construction in physics, yet many learners understand them as exact replicas (ERs) of reality rather than as functional, theory-guided tools. This study examines whether the fidelities model of conceptual development, which distinguishes between fidelity of gestalt (FG) and functional fidelity (FF), also applies to learners' general, domain-independent understanding of models. To this end, established "ER" items were systematically revised to separate visual and functional aspects of model understanding. A digital cross-sectional survey was conducted with N = 231 students in grade 7 to grade 10 at a German secondary school. Data was analyzed using confirmatory factor analysis, descriptive statistics, and classification based on weighted FG and FF scores. The results support the proposed two-factor structure of general model understanding, distinguishing between FG, which emphasizes visual similarity, and FF, which emphasizes functional representation. Typological analysis identified the dual type as most frequent (45.02%), followed by the undeveloped (25.10%), functional (21.65%), and architectural model (8.23%) types. Overall, the findings validate a domain-independent instrument for assessing general model understanding and suggest that model competence should be treated as a multidimensional construct in science education.

Keywords: mental models, model understanding, fidelity of gestalt, functional fidelity

INTRODUCTION

The acquisition of knowledge in physics is hardly conceivable without the use of models. They not only serve as a means of communication for abstract concepts but are also essential tools for making theoretical predictions and testing hypotheses (Treagust et al., 2002). In educational practice, however, a significant gap regularly emerges between teachers' understanding of models and students' intuitive, often concrete conceptions. For many students, models are not perceived as theory-driven, functional constructs. As classical studies by Treagust et al. (2002) have shown, learners tend to mistakenly interpret models as exact physical replicas of reality. However, anyone who ontologizes scientific representations in this way will inevitably encounter insurmountable cognitive barriers, particularly in more abstract domains (Grosslight et al., 1991; Treagust et al., 2002). The identification and classification of these mental conceptions is therefore indispensable in order to support learners in developing an appropriate level of abstraction (Grosslight et al., 1991).

In previous subject-didactic research, understanding of models was initially categorized qualitatively (Grosslight et al., 1991) and later quantitatively (Treagust et al., 2002) into measurable dimensions. Building on these findings, Ubben and Bitzenbauer (2022) developed the fidelity model of conceptual development (FMCD) (cf. Ubben & Heusler, 2021). This framework postulates that model understanding is based on two dimensions: fidelity of gestalt (FG), which focuses on visual similarity, and functional fidelity (FF), which emphasizes the functional aspects of phenomena. Previous studies have already successfully confirmed this structure for specific topics such as quantum physics (Bitzenbauer & Ubben, 2025; Ubben & Bitzenbauer, 2023). However, it has remained unclear to what extent it can also represent general, domain-independent model knowledge.

✉ hw39syva@studserv.uni-leipzig.de ✉ joaquin.veith@uni-leipzig.de ✉ malte.ubben@uni-leipzig.de **(*Correspondence)**

Contribution to the literature

- This study extends the FMCD beyond domain-specific contexts by demonstrating that the two-dimensional structure (FG and FF) also applies to learners' general, domain-independent model understanding.
- It introduces a methodological refinement by systematically decomposing "ER" items into separate visual and functional components, enabling a more precise and psychometrically robust measurement of model understanding.
- It provides new empirical evidence that general model metaknowledge constitutes a distinct cognitive challenge, highlighted by the substantial presence of a non-developed learner type, thereby offering a novel perspective on learning barriers in science education.

The present study aims to close this research gap. Its methodological core lies in the systematic decomposition and re-operationalization of the classical "exact replica (ER)" items by Treagust et al. (2002), as these originally conflated visual and functional aspects. Although the FMCD originated from physics education research, the present study deliberately investigates whether the underlying cognitive dimensions extend beyond specific physics contexts to general model understanding. The goal is to validate a discriminant, domain-independent instrument. Previous studies have consistently validated the FMCD within specific scientific domains. The present study extends this line of research by investigating whether the proposed two-dimensional structure represents a general characteristic of learners' model understanding rather than a domain-specific phenomenon. To achieve this, established "ER" items were systematically reformulated to separate visual (FG) and functional (FF) aspects, resulting in the first domain-independent instrument designed to validate the FMCD beyond individual scientific contexts. Using confirmatory factor analysis (CFA), the study thus empirically examines whether the two-dimensional architecture of the FMCD can also be demonstrated in learners' general metacognitive understanding of models, and which cognitive types can be derived from this in lower secondary education.

THEORETICAL BACKGROUND

Understanding models is a central component of scientific practice (cf. Gilbert, 2004; Schwarz et al., 2009). A well-founded comprehension of how these tools function (and of their limitations) constitutes a key pillar of basic scientific literacy, i.e., the ability to understand and critically use scientific concepts and practices (Gräber et al., 2002; OECD, 2016) and is frequently discussed in the literature under the concept of the nature of science (NOS), referring to learners' understanding of how scientific knowledge is generated, justified, and revised (Lederman, 2007).

In the present study, scientific models (or simpler: "models") refer to the external representational tools used in science, whereas "mental models" denote learners' internal representations of these scientific models and their function. Underlying the use of every scientific model is thus a crucial question: how do learners think about scientific models? While teachers employ scientific models as functional constructs and tools for generating knowledge, learners' cognitive perspective on these models often remains a black box (Treagust et al., 2002). Investigating mental models is therefore not a purely statistical end in itself, but a necessary prerequisite for uncovering learners' often implicit conceptions and guiding them toward a reflective and scientifically appropriate understanding of scientific models. An early qualitative milestone in addressing this question was the interview study by Grosslight et al. (1991). The authors demonstrated that understanding of (scientific) models develops across three hierarchical levels: At the first level, learners view models simply as reduced, real copies of reality. At the second level, they recognize the purpose of a model but still remain strongly attached to its superficial similarity to the original. Only at the third level (the expert level) does a truly reflective understanding emerge. Here, the model is seen for what it is: a theoretical, abstract construct used to test hypotheses and make predictions.

Building on this qualitative work, the study by Treagust et al. (2002) enabled the first quantitative assessment of these competencies through the development of the students' understanding of models in science (SUMS) questionnaire. This instrument made learners' latent conceptions measurable for the first time by structuring model understanding into five dimensions, including ERs, models as explanatory tools, multiple models, changing models, and uses of scientific models.

In terms of data collection, the SUMS questionnaire uses a five-point Likert scale. The statistical analyses based on this instrument not only validated the theoretical model levels across large samples but also revealed a key issue: although many learners do recognize the explanatory function of scientific models, they struggle significantly to internalize their provisional and theory-driven nature. In particular, the notion of a scientific model as an ER or copy proved to be a major obstacle on the path toward a genuine expert-level understanding.

These findings inevitably raise the question of how strongly this general epistemological understanding of scientific models is connected to subject-specific knowledge in concrete domains, and to what extent an inadequate understanding can explain common learning barriers. This can be illustrated using the Bohr model, where a fixation on the idea of an "ER" can severely hinder the learning process: if representational conventions such as colored spheres or fixed orbits are ontologized as material reality rather than understood as functional auxiliary constructs, the cognitive pathway toward more abstract physical theories remains blocked (Ubben, 2020).

In his investigation into the typology of mental model understanding, Ubben (2020) addressed this issue using the example of quantum physics. He adapted selected SUMS items by David and combined them with subject-specific tasks on the atomic shell based on the work of Müller and Wiesner (2002). Contrary to the expectation of two separate clusters, general scientific model understanding and concrete physical conceptions could not be cleanly separated statistically. Instead, a deep structural interconnection between the items emerged.

This suggests that learners' domain-specific concepts and their general understanding of science (particularly in highly abstract fields such as quantum physics) may be traceable to a common underlying cause. This idea was further developed (Ubben, 2020) and led to the framework of FG and FF. This framework offers a new perspective on how learners evaluate models: are they primarily guided by visual, intuitive structure (FG), or do they focus on mathematical-functional predictive power (FF)?

FMCD

The starting point for FMCD was the pioneering study conducted in the context of atomic models, in which the model was successfully validated for the first time (Ubben & Heusler, 2021). Building on these relatively intuitive model conceptions, research interest subsequently expanded to additional subject areas. The framework was therefore not only applied to investigate students' conceptions of the photon (Ubben & Bitzenbauer, 2022), but also used to explore further physical phenomena, for example, through qualitative analyses of mental models related to black holes (Ubben et al., 2022) and the Higgs boson (Bitzenbauer & Fösel, 2025).

Based on this two-dimensional structure, empirical findings on students' conceptions can be represented in a matrix of four archetypal types (see Figure 1):

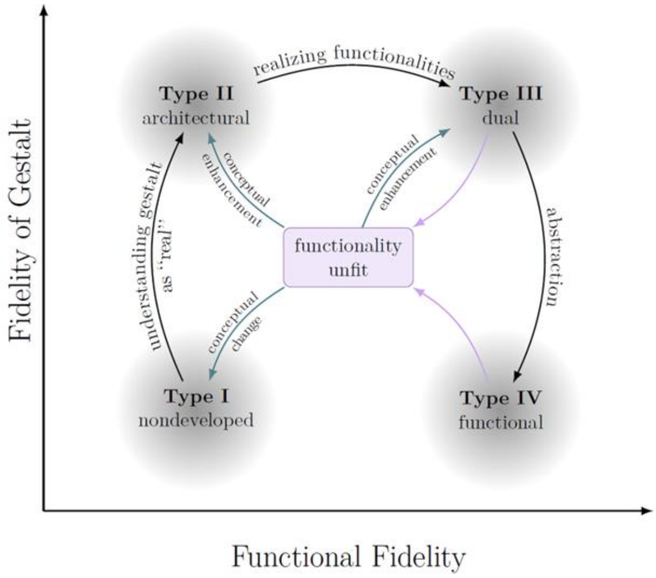

Figure 1. FMCD: Starting from the perception of a particular form as "real," an abstract understanding of scientific models is developed through the exploration of their functionality (this understanding evolves through an iterative process: if a given model proves to be no longer applicable, it must be modified or further developed) (Bitzenbauer & Ubben, 2025, p. 17).

1. **The non-developed type (low FG, low FF):** At this stage, model understanding is largely undifferentiated. Learners lack a metacognitive grasp of both the representational and functional levels of a model. Their knowledge is often limited to the recall of isolated, declarative facts, without establishing connections between the model structure and the phenomenon. In this quadrant, the model is perceived neither as a tool for gaining insight nor as an adequate representation (cf. Ubben, 2020; Ubben et al., 2023). Example (photon): Learners of this type have no prior knowledge of the concept of the photon; accordingly, their conception remains undeveloped (Ubben & Bitzenbauer, 2022, p. 13).
2. **The architectural model type (high FG, low FF):** This type is characterized by a strong focus on appearance/ gestalt. The quality of a model is primarily judged based on its visual similarity to the original. Learners in this quadrant perceive the model as a purely illustrative tool, whose value lies in aesthetic or structural resemblance, while its abstract, explanatory character remains largely unrecognized (cf. Ubben, 2020; Ubben et al., 2023.). Example (photon): Learners of this type describe a photon as a localized "billiard ball that carries energy and momentum" (Ayene et al., 2011), or imagine it as a small, spherical particle moving along a wave-like trajectory (see Krijtenburg-Lewerissa et al., 2017; Ubben & Bitzenbauer, 2022, p. 13).
3. **The dual type (high FG, high FF):** In this type, functionality and gestalt merge into a single cognitive construct. Learners attribute strong predictive power to the model but justify this by assuming that the model is a "scaled up or down

replica" (Bitzenbauer & Ubben, 2025, p. 13) of reality. The frequently observed high intercorrelation between FG and FF arises from the belief that a model can only be functional if it represents reality as accurately and in as much detail as possible. An abstract understanding of the necessary reduction and idealization inherent in scientific models is not yet fully developed (cf. Ubben, 2020; Ubben et al., 2023). Example (photon): Learners initially describe photons in a highly abstract way as energy quanta of light. However, when asked to explain specific experiments, they revert to a naïve particle-based conception (cf. Bitzenbauer & Meyn, 2022).

4. **The functional type (low FG, high FF):** This type most closely corresponds to an expert-level understanding. Here, a cognitive decoupling from external form takes place. The concrete visual appearance becomes less relevant, as the model's form "is only seen as a carrier […] for the processes encapsulated in the idea" (Bitzenbauer & Ubben, 2025, p. 13). Thus, the external representation is perceived as arbitrary or necessarily abstract, as long as it fulfills its functional purpose. Learners of this type are capable of appropriately using even highly abstract models without macroscopic analogues (cf. Ubben, 2020; Ubben et al., 2023).). Example (photon): Learners consistently describe photons in a highly abstract manner as energy quanta of light, without reverting to naïve, object-based conceptions (cf. Ubben & Bitzenbauer, 2022, p. 13).

That the FMCD is also suitable for diagnostic purposes in school practice is demonstrated by more recent replication studies, such as those in the field of linear polarization of light (Schmid et al., 2026). In this context, the framework was validated as a robust instrument for identifying specific learner types in secondary education. Additionally, the scope of the FMCD is no longer limited to the domain of physics; its successful transfer to mathematics. For instance, it was successfully applied in the investigation of students' conceptions of angles (Veith et al., 2025). This highlights the potential universal applicability of the cognitive dimensions of FG and FF.

This series of replications across a wide range of content areas and disciplines attests to the strong theoretical robustness of the FMCD. However, while these studies provide valuable insights into domain-specific topics, the question remains to what extent these cognitive structures represent general, domain-independent features of model understanding. To move beyond this content specificity and develop a meta-perspective on cognitive abstraction processes, further methodological refinement is required.

The present research therefore focuses on refining and generalizing test items originally based on the work of Treagust et al. (2002), which had already served as a foundation in the seminal study by Ubben and Heusler (2021). The aim is to revise these instruments in such a way that they yield robust conclusions about the nature of mental models regarding FMCD that remain valid independently of specific physical phenomena.

A central focus lies on the category of ER items (models as ERs), as originally defined by Treagust et al. (2002). These items were initially designed to assess the extent to which learners perceive a scientific model as an exact reproduction of reality. However, closer examination reveals that the original formulations partially conflated aspects of visual similarity with functional explanatory power at a conceptual level.

In order to do justice to the theoretically grounded two-dimensionality of model understanding, a systematic separation of these constructs was essential. The original items were therefore reformulated into distinct statements for the present study. This approach enables a clear and isolated operationalization of FG and FF.

Building on the research questions proposed by Schmid et al. (2026, p. 2), these are adapted here to address learners' general conceptions of scientific models:

1. Does a two-factor structure characterized by FG and FF, as proposed in the FMCD, also hold for learners' general understanding of scientific models?
2. In accordance with the FMCD, which types of understanding can be identified among learners with respect to their general understanding of mental models?

METHODOLOGY

Design of the Study

The study was conducted as a digital cross-sectional survey using the software LimeSurvey. Data collection took place at a secondary school (gymnasium) in Germany and included a total of N = 231 students from grades 7 to 10. A detailed description of the sample is provided in the sub-section on sample characteristics. The cross-sectional survey design was chosen because the study aims to validate the latent factor structure of the proposed instrument rather than to investigate instructional effects or developmental changes.

To ensure valid responses to the items and to establish a shared understanding of the concept of scientific models, participants were given a standardized explanatory text prior to completing the survey. A total of 30 items were used to assess mental models, all answered on a five-point Likert scale (1 = strongly

Table 1. Original item phrasings from Treagust et al. (2002) of their category ER and updated ones considering FF and FG representation

Original item	New code	New item
A model needs to be close to the real thing by being very exact, so nobody can disprove it.	F2	A model must be close to reality in terms of how it functions.
A model should be an ER.	F7	A model should replicate the function of a real object/phenomenon exactly.
	G7	A model should replicate the appearance of a real object/phenomenon exactly.
A model needs to be close to the real thing.	F8	A model must be close to reality with regard to how it works.
	G8	A model must be close to reality in terms of its appearance.
A model needs to be close to the real thing by being very exact in every way except for size.	F10	A model must come very close to reality by being very precise in every functional aspect.
	G10	A model must come very close to reality by being very precise in all visual aspects.
A model shows what the real thing does and what it looks like.	F12	A model shows what the real object/phenomenon does.
	G12	A model shows what the real object/phenomenon looks like.

disagree, 2 = rather disagree, 3 = partly/neutral, 4 = rather agree, 5 = strongly agree). The study was designed as a purely observational research design; a targeted instructional intervention was explicitly not part of the investigation.

Item-Design and Instrumentation

The development of valid test instruments within the framework of the FMCD has, in previous research, relied heavily on statistical validation with CFA. This method has been used to empirically confirm the theoretically postulated distinction between FG and FF. Corresponding instruments have already been successfully validated across various domains in physics and mathematics, including previously mentioned examples such as light polarization (Schmid et al., 2026), photon concepts in quantum optics (Bitzenbauer & Ubben, 2025), and geometric conceptions of angles (Veith et al., 2025).

In contrast to these studies, the present investigation does not introduce entirely new items. Instead, it focuses on a systematic reanalysis and modification of the established items developed by Treagust et al. (2002). A critical feature of the original Treagust items is their often hybrid structure: many statements simultaneously address both functional and gestalt-related aspects of a model within a single item. This complicates a clear separation of the cognitive dimensions in the sense of the FMCD.

The core of the instrument design used here therefore lies in the deliberate decomposition of these (partially) combined statements (see **Table 1**). As originally defined: "The ER scale refers to students' perceptions of how close a model is to the real thing" (Treagust et al., 2002, p. 359).

Through this procedure, the study aims to demonstrate that the cognitive structures described in the FMCD are already implicitly embedded in the foundational work of Treagust et al. (2002) and Grosslight et al. (1991). By refining item formulations, these structures can be operationalized in a more precise and domain-independent manner, thereby extending and sharpening the contributions of prior research.

Sample

To explore mental models about scientific models, a cross-sectional research design was employed. The inclusion of grades 7 to 10 was intended to capture a sufficiently large and cognitively heterogeneous sample representing the full spectrum of lower secondary education ("Sekundarstufe I – Gymnasium").

Accordingly, data collection took place at a single secondary school, with the survey being directly integrated into regular physics lessons. The selection of this site was primarily justified by its characteristics as a typical Saxon gymnasium, ensuring that the participants reflect the curricular standards and cognitive demands associated with this type of school. In addition, the school's size and multiple class streams provided the necessary infrastructure to collect a sufficiently large sample (N = 246) under controlled and comparable classroom conditions.

Prior to test administration, students did not receive any instruction or targeted intervention on model construction beyond the standard curriculum. The instrument was administered digitally via LimeSurvey. To ensure the validity of the statistical analyses, only fully completed datasets were included in the final evaluation (N = 246).

Participant recruitment was based on voluntary participation in the sense of non-probabilistic convenience sampling. In total, the questionnaire was distributed to N = 246 students from grades 7 to 11. Since only 15 participants were enrolled in upper secondary education (Sekundarstufe II), this subgroup was too small to permit meaningful statistical analyses or comparisons with lower secondary students. To ensure a homogeneous sample aligned with the study's

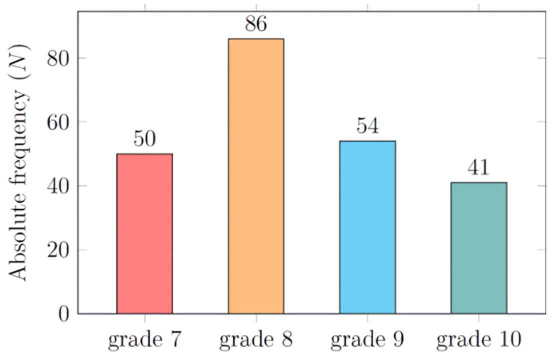

Figure 2. Distribution of queried students regarding years (N = 231) (the authors' on elaboration)

Table 2. Results of confirmatory factor analysis regarding underlying model (comparing one-factor to two-factor model)

Criterion	Threshold	One-factor model	Two-factor model
χ^2	-	87.5	31.2
df	-	19	18
p-value	-	< .001	0.027
CFI	≥ 0.95	0.865	0.973*
TLI	≥ 0.95	0.801	0.958*
SRMR	≤ 0.08	0.066*	0.035*
RMSEA	≤ 0.05	0.113	0.056
AIC	-	5,166	5,112
BIC	-	5,253	5,202

Note. * indicates passing of the criterion

research questions, subsequent analyses were therefore restricted to students in grades 7-10 (Sekundarstufe I; N = 231) (see **Figure 2**). The age distribution of the investigated cohort ranged from 11 to 17 years, thus corresponding to the typical age range of the respective grade levels.

With regard to the sociocultural and institutional context, it should be noted that the secondary school of the study is not located in a socially disadvantaged area and has a relatively low proportion of students with language-related difficulties. Furthermore, since the school does not have a specialized focus or profile in the natural sciences, it can be assumed that the participants possess a level of prior knowledge in science that is typical for general academic-track secondary schools.

DATA ANALYSIS

To address research question 1, a CFA was conducted to examine whether learners' general model understanding is best represented by the proposed two-factor structure consisting of FG and FF. This procedure for establishing construct validity followed the methodological approach of Bitzenbauer and Ubben (2025).

The global model fit was evaluated using a set of standard indices, including the χ^2 goodness-of-fit test, the comparative fit index (CFI), the Tucker-Lewis index (TLI), the root mean square error of approximation (RMSEA), and the standardized root mean residual (SRMR). While the CFI indicates the relative improvement in model fit compared to a baseline model, the TLI additionally accounts for the number of degrees of freedom and penalizes unnecessary model complexity, making it a particularly stringent indicator of model parsimony.

For the interpretation of model fit, the threshold values recommended by Schermelleh-Engel et al. (2003) and Cho et al. (2020) were applied; these, together with the values for the different models, are presented in **Table 2**.

To systematically evaluate both model fit and model complexity, a dual approach was adopted that combines information-theoretic indices with statistical comparisons. Specifically, the Akaike information criterion (AIC) and the Bayesian information criterion (BIC) were calculated to assess competing model specifications. These criteria penalize overparameterization and thus favor more parsimonious models over unnecessarily complex ones (Vrieze, 2012). While the AIC focuses on optimizing predictive accuracy, the BIC imposes a stronger penalty on model complexity (particularly relevant for larger sample sizes) and therefore tends to favor more concise models (Chakrabarti & Ghosh, 2011).

In addition, a one-factor model was estimated as a comparative baseline in order to account for the originally proposed unidimensional structure of model understanding by Treagust et al. (2002).

To further complement these metrics, a formal inferential statistical step was incorporated through the use of likelihood ratio tests (LRTs). These tests evaluate whether a more complex model provides a statistically significant improvement in fit compared to a simpler baseline model. By jointly considering AIC, BIC, and LRTs, a balance is achieved between predictive performance, model parsimony, and statistical significance, thereby providing a robust foundation for final model selection (Vrieze, 2012).

To evaluate local model fit at the level of individual indicators, standardized factor loadings (λ) were analyzed, along with the associated error variances (calculated as $1 - \lambda^2$) and the squared λ (λ^2) as measures of item reliability. Cross-loadings were not permitted in the model specification. Items with standardized loadings below 0.30 were excluded from the final model in accordance with the recommendation of Kline (2000).

Reliability at the level of latent constructs was assessed using McDonald's omega (ω), which provides a more appropriate estimate than Cronbach's alpha when λ are not uniform across indicators. A threshold of $\omega > 0.70$ was defined as an indicator of acceptable internal consistency for each scale (cf. McDonald, 1999).

In addition, the average variance extracted (AVE) was calculated for each latent construct to assess convergent and discriminant validity. AVE quantifies the proportion of variance in the observed indicators that can be attributed to the underlying construct, relative to the proportion due to measurement error (cf. Brown, 2015). Following the criterion proposed Fornell and Larcker (1981), discriminant validity can be considered adequate if the AVE of a construct exceeds its squared correlations with any other construct in the model.

To complement the CFA with a broader perspective on the data, key descriptive statistics were also reported, including response distributions, means (Ms), medians, and standard deviations (SDs) for each indicator.

To address research question 2, basic descriptive statistics for all indicators were first examined, including response distributions, Ms, SDs, and median values. Subsequently, the results of the CFA were used to analyze learners' mental models by assigning them to the typology described in the methodology section.

This classification was based on their composite FG and FF scores, allowing the distribution across the four model types to be represented. For this purpose, two composite metrics were derived to represent the respective latent constructs based on the λ: the FF score for the FF factor and the FG score for the FG factor. These scores were calculated as weighted Ms of the observed indicators F_i and G_i, where the weights correspond to the standardized λ_i:

$$FF\,score = \frac{\sum_{i=1}^{n} \lambda_i \cdot F_i}{\sum_{i=1}^{n} \lambda_i}, \tag{1}$$

and

$$FG\,score = \frac{\sum_{i=1}^{n} \lambda_i \cdot G_i}{\sum_{i=1}^{n} \lambda_i}. \tag{2}$$

This weighting method ensures that indicators with higher λ (i.e., those more strongly associated with the latent factor) contribute more substantially to the respective composite score.

RESULTS

Results concerning Research Question 1: Global Model Fit

Table 2 presents a comparison of the fit indices for the two examined model variants. The statistical evaluation clearly shows that reducing the structure to a single factor is associated with a significant deterioration in model fit. This is particularly evident in the lower CFI value but is also clearly reflected in the reduced TLI value.

This finding is further supported by the SRMR and RMSEA values, which exceed acceptable threshold levels in the one-factor model, thereby indicating an inadequate representation of the empirical data. In direct contrast, the two-factor model proves to be clearly superior. It not only achieves a lower χ^2 statistic but also meets all established criteria for global model fit reported in the literature.

Table 3. Results of the CFA for the two-factor model, including item estimates (factor loadings, standard errors, error variance) as well as reliability indices (indicator reliability, factor reliability ω, and average variance extracted, AVE)

Factor	Item	λ (SE)	ε	λ^2	ω	AVE
FF	F2	0.604 (0.078)	0.635	0.365	0.73	0.41
	F7	0.699 (0.083)	0.511	0.489		
	F8	0.658 (0.071)	0.567	0.433		
	F10	0.591 (0.066)	0.651	0.349		
FG	G7	0.776 (0.070)	0.398	0.602	0.83	0.55
	G8	0.817 (0.074)	0.333	0.667		
	G10	0.814 (0.066)	0.337	0.663		
	G12	0.531 (0.080)	0.718	0.282		

Note: SE: Standard error; ε: Error variance (1 - λ^2); λ^2: Indicator reliability; & ω: Factor reliability

Additionally, the LRT confirmed that the two-factor model provides a significantly better fit to the empirical data ($p < 0.001$). This superiority is also evident when considering the information criteria AIC and BIC. The lower values obtained for the two-factor model indicate a more balanced relationship between explanatory power and model complexity (parsimony).

Taken together, these findings indicate that learners' responses are better explained by distinguishing between FG and FF than by assuming a single, undifferentiated dimension of model understanding. Thus, the statistical analyses provide empirical support for the two-dimensional structure proposed by the FMCD.

Results concerning Research Question 1: Local Model Fit

Next, the local fit indices at the level of individual indicators are considered (see **Table 3**). For the FF factor, the standardized λ fall within a moderate to good range between 0.591 and 0.699. All items therefore clearly exceed the critical threshold of 0.30 (cf. Kline, 2000) and exhibit substantial proportions of systematic variance attributable to the latent construct. The factor reliability for FF, with ω = 0.73, can be considered good and indicates sufficient internal consistency of the scale (cf. McDonald, 1999). The AVE, at 0.41, falls below the often conservatively required threshold of 0.50; however, given the solid factor reliability, it can still be regarded as acceptable within the context of exploratory scale development (cf. Fornell & Larcker, 1981).

With respect to FG, predominantly very strong loadings are observed, with two indicators reaching values of λ > 0.810. Only item G12 shows a somewhat lower loading of 0.531, but it still contributes significantly to the definition of the factor. One possible explanation is that, unlike the remaining FG items, G12 does not formulate visual similarity as a normative

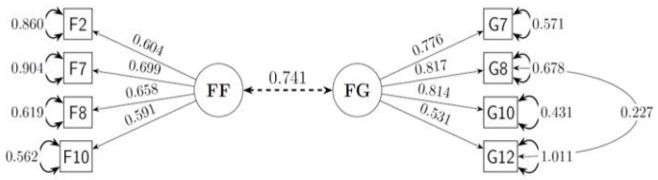

Figure 3. Path diagram of the two-factor model with standardized factor loadings, error variances, and the residual covariance between G8 and G12 due to similar phrasing (the authors' on elaboration)

Indikator	Mean ± SD	Mean	SD	Mdn	+ (%)	- (%)
F2		3.33	1.11	3	47.19	26.41
F7		3.12	1.18	3	40.69	33.77
F8		3.75	1.03	4	66.67	13.42
F10		3.43	0.96	4	53.25	19.91
FF Score		3.40	0.78	3.47	–	–
G7		2.89	1.09	3	29.87	39.39
G8		3.23	1.16	3	46.75	30.74
G10		3.16	1.05	3	39.39	29.44
G12		3.21	1.14	3	41.99	28.14
FG Score		3.12	0.87	3.18	–	–

Figure 4. Descriptive statistics including graphical representation of M and SD (in addition, the graph shows the percentage of explicit agreement (+, values 4 and 5) and disagreement (-, values 1 and 2) for all items; neutral responses (value 3) are not considered) (Five-point Likert scale: 1 = strongly disagree, 2 = rather disagree, 3 = partly/neutral, 4 = rather agree, 5 = strongly agree) (the authors' on elaboration)

requirement for a good scientific model but merely states that models show what real objects or phenomena look like. Since many scientific models indeed possess a visual representational component, agreement with this item may also occur among learners who otherwise emphasize the functional role of models, thereby reducing its discriminatory power within the FG construct. The indicator reliabilities (λ^2) for G7, G8, and G10 all exceed 0.600, meaning that more than 60% of the variance in these items can be explained by the latent construct FG. The scale demonstrates high factor reliability, with $\omega = 0.83$. Furthermore, the AVE of 0.55 exceeds the threshold of 0.50, thereby empirically supporting the convergent validity of the FG factor. The residual covariance between G8 and G12 was retained because both items refer explicitly to the visual appearance of models ("appearance" and "looks like"), giving rise to additional shared method variance due to similar wording beyond the common FG construct. In summary, the selected items of the short scale reliably represent the theoretical dimensions FG and FF. This indicates that the revised item set captures both dimensions of the FMCD with sufficient reliability and supports their use for assessing general model understanding.

Beyond internal consistency, the empirical distinguishability of latent constructs is a key prerequisite for instrument validity. The correlation between the factors FG and FF in the present sample is $r = 0.741$. To assess discriminant validity, the criterion proposed by Fornell and Larcker (1981) was applied, which requires that the AVE of each construct exceed the squared correlation with other factors. With a squared correlation of $r^2 \approx 0.55$, the following relation emerges for the present data: $0.41 = AVE_{FF} < 0.55 \approx r^2 > 0.55 = AVE_{FG}$.

Thus, while the FG factor satisfies the criterion, the value for FF falls below the squared correlation. Despite this high conceptual proximity between the dimensions, the significantly superior global model fit of the two-factor solution (see **Table 2**), together with the theoretical assumptions of the FMCD, justifies retaining the separate factors.

The results suggest that FG and FF are highly associated but nonetheless represent distinct facets of model understanding that can be differentiated by the instrument. The resulting path diagram is illustrated in **Figure 3**.

In summary, the findings of the CFA support the assumption that learners' mental models describing the scientific models are underpinned by a two-dimensional structure. The successful modeling demonstrates that refining the original ER items enables a clear and discriminant measurement of cognitive model conceptions. The complete questionnaire initially comprised 30 items addressing different aspects of model understanding. However, the present study focuses exclusively on the ER category proposed by Treagust et al. (2002), since these items were systematically reformulated to distinguish between FG and FF. **Table 1** therefore reports only the revised ER items that formed the basis of the CFA presented in this paper.

Results Concerning Research Question 2

To address the second research question, the distribution of mental models within the cohort was analyzed based on descriptive statistics (**Figure 4**) and the resulting classification within the quadrant model (**Figure 5**).

The analysis of individual items reveals a tendency toward higher agreement in the domain of FF. The M values of the FF scale range from M = 3.12 (SD = 1.18) for item F7 to M = 3.75 (SD = 1.03) for item F8. For two FF items, the median is 4, indicating a clear tendency toward explicit agreement. For example, 66.67% of respondents show a positive evaluation of item F8. Overall, the functional dimension yields a score of M = 3.40 (SD = 0.78).

In contrast, the evaluations of FG are more reserved. The M values for FG items range from M = 2.89 (SD = 1.09) for item G7 to M = 3.23 (SD = 1.16) for item G8, with medians consistently at 3. This suggests that learners are

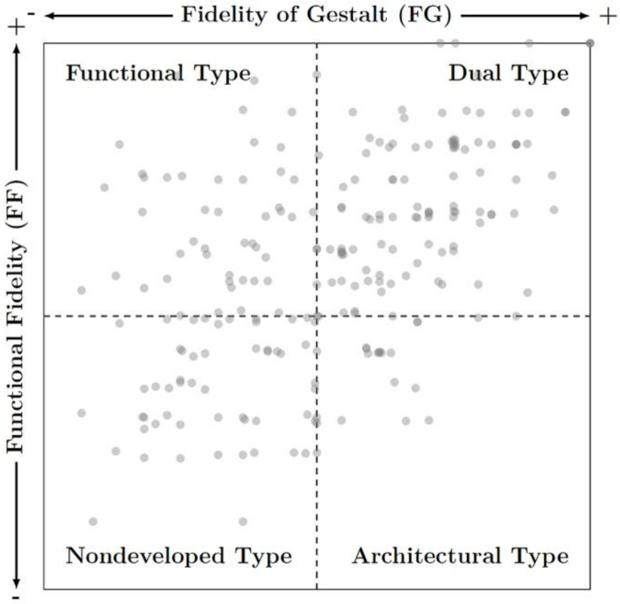

Figure 5. Distribution of participants (weighted scores) across the quadrants of the four types of mental models (the authors' on elaboration)

more skeptical toward, or more neutral about, the external representational form and visual aspects of models compared to their functional aspects. Accordingly, the FG score (M = 3.12, SD = 0.87) is lower than that of the functional dimension.

The distribution of individual weighted scores within the typology framework (**Figure 5**) reflects the cognitive diversity of the cohort. The dual type constitutes the largest group at 45.02%, followed by the undeveloped type at 25.10%. The functional type accounts for 21.65%, while only 8.23% of participants are classified as the architectural type.

Thus, the results indicate that nearly two-thirds of learners possess a consolidated understanding of model functionality (sum of dual and functional types). At the same time, the substantial proportion of the undeveloped type highlights that for approximately one quarter of participants, neither dimension of model fidelity nor represents a stable conceptual framework. Overall, the distribution suggests that although many learners already recognize the functional role of models, a substantial proportion still exhibit conceptions that are either strongly linked to visual similarity or remain insufficiently differentiated.

DISCUSSION

Discussion Regarding Research Question 1

The empirical findings of the CFA support the assumption that learners' general conceptions of models can be represented by a two-dimensional structure. The comparison of the estimated models highlights the superiority of this theoretical distinction: whereas the one-dimensional solution fails to adequately represent the data structure (CFI = 0.865 < 0.95; TLI = 0.801 < 0.95; RMSEA = 0.113 > 0.06), the two-factor model achieves excellent fit with a CFI of 0.973 and a TLI of 0.958. The fact that the absolute error indices (SRMR = 0.035; RMSEA = 0.056) also fall well below critical thresholds demonstrates the robustness of this model for the general assessment of model knowledge.

The statistical relevance of the second factor is further reflected in the development of the information criteria (AIC and BIC) as well as in the significance level of the χ^2 test. The increase in the χ^2 test p-value from < 0.001 in the one-factor model to 0.027 in the two-factor model, together with the improved fit indices, indicates that differentiating between FG and FF substantially reduces the discrepancy between the theoretical model and the empirical data. This suggests that learners are already capable at a general level of distinguishing between a model's visual resemblance to reality and its functional purpose, even in the absence of a specific physical context.

A closer inspection of local model fit, however, reveals a methodological peculiarity. With an intercorrelation of r = 0.741, the two dimensions exhibit a considerably stronger relationship than is often observed in domain-specific contexts (e.g., in quantum physics, see Bitzenbauer and Ubben, 2025)). This high correlation leads to the Fornell-Larcker criterion not being fully satisfied for the FF factor (AVE = 0.41 < $r^2 \approx$ 0.55). Substantively, however, this can be explained by the fact that, when assessing general model conceptions, different aspects of models are often still cognitively intertwined. Many learners appear to regard high visual similarity (FG) as a prerequisite for high functionality (FF). The findings also contribute to ongoing discussions regarding the relationship between visual representations and scientific reasoning. Rather than viewing representational fidelity and functional understanding as competing perspectives, the FMCD suggests that many learners initially integrate both dimensions. Only later does functional reasoning become increasingly independent of visual similarity, consistent with the developmental assumptions underlying the model. In addition to this cognitive interdependence, the statistical relationship is also strongly influenced by the distribution of model types within the sample. The high proportion of participants classified as the dual type (45.02%), who by definition exhibit high levels in both dimensions, increases the shared variance between the factors and is thus reflected strongly in the elevated intercorrelation.

From the perspective of the FMCD, these findings suggest that FG and FF are not merely characteristics of learners' understanding within individual scientific domains but represent more general cognitive dimensions underlying model thinking. Previous validations of the FMCD focused on specific contexts such as quantum physics or light polarization. The present findings extend this theoretical framework by demonstrating that the same two-dimensional structure can also be observed when learners reason about scientific models in general, independent of a particular disciplinary context. This supports the assumption that the FMCD captures fundamental aspects of model cognition rather than domain-specific knowledge alone.

Discussion Regarding Research Question 2

The distribution of participants within the FMCD framework provides a differentiated picture of general model understanding in lower secondary education (Sekundarstufe I). A key finding is the dominance of the dual type, which accounts for 45.02% of the sample; nearly half of all participants. This strong concentration in the first quadrant is not merely a descriptive result but also forms the statistical basis for the high intercorrelation between the factors ($r = 0.741$) discussed in the previous section.

Since a large proportion of learners attribute both high visual fidelity (FG) and high FF to models, their evaluations on both dimensions tend to cluster synchronously in the upper range. This close interconnection distinguishes the present study from investigations of more abstract topics, such as quantum physics (cf. Bitzenbauer and Ubben, 2025), where Gestalt and function are often cognitively more clearly separated.

A distinctive feature of the present findings, compared to prior FMCD research (e.g., Bitzenbauer & Ubben, 2025; Schmid et al., 2026), is the prominent representation of the undeveloped type (25.10%). In earlier studies, this quadrant was rarely or not at all identified, pointing to an important methodological difference: whereas previous research consistently focused on domain-specific topics (such as quantum physics or polarization) the instrument used here targets general meta-knowledge about models without a subject-specific context.

It can be assumed that this higher level of abstraction presents a greater cognitive challenge for a significant portion of learners. A concrete physical topic often serves as a conceptual anchor that makes the functional role of a model (FF) in the learning process almost intuitively accessible. In the absence of such context, evaluating general statements about models requires purely metacognitive reflection, detached from direct application.

Without this “anchoring function,” many lower secondary students appear to revert to a pre-scientific understanding, in which models are primarily conceived as concrete representations of reality. The fact that the undeveloped type becomes visible to this extent for the first time thus provides an important indication of the domain specificity of model competence. It suggests that the ability to reflect on models is often more developed within specific subject-matter contexts than at a purely abstract, theoretical level.

Within the FMCD, the identified learner types represent different ways of evaluating scientific models rather than fixed stages of development. The predominance of the dual type therefore suggests that many learners continue to associate explanatory quality with visual realism. From the perspective of the FMCD, this represents an important transitional state: learners recognize that models should explain phenomena, yet they still judge explanatory power largely through visual resemblance. The comparatively large proportion of non-developed learners further indicates that abstract reflection about models remains a challenging metacognitive task in lower secondary education.

Limitations

The present study is subject to several methodological limitations that should be considered when interpreting the results. The chosen cross-sectional design primarily provides a snapshot of learners’ model understanding and does not allow for direct conclusions about developmental trajectories or causal relationships (cf. Döring & Bortz, 2016). In addition, the data collection is based on self-reports using Likert scales, meaning that common biases, such as socially desirable responding, cannot be entirely ruled out (Podsakoff et al., 2003).

At the same time, the use of a short scale represents a deliberate trade-off between efficiency and conceptual differentiation, which nevertheless appears appropriate given the clearly identifiable factor structure. The characteristics of the sample also limit the generalizability of the findings. The study is based on a convenience sample from a single school and focuses on lower secondary education, meaning that transferability to other school types or age groups is only possible to a limited extent (Cohen et al., 2017). However, the sample reflects the essential features of a typical school learning environment, allowing the results to be interpreted as a meaningful contribution to understanding general model competence in educational contexts.

At the level of the measurement instrument, it is important to note that the items are derived from established scales and were systematically refined in line with the FF-FG framework (cf. Ubben, 2020). This approach enables a more precise separation of the theoretical dimensions, although further validation in different contexts is required. The absence of qualitative methods, such as interviews, limits insights into individual interpretative processes. At the same time, the consistent factor structure suggests that the intended constructs were captured adequately overall, even though the high level of abstraction of the items likely posed an additional challenge for some learners.

From a theoretical and statistical perspective, a comparatively high intercorrelation between FG and FF is observed. However, this can be plausibly interpreted and does not primarily represent a weakness of the model. Rather, it can largely be explained by the distribution of model types: the dominance of the dual type leads to both dimensions being highly expressed simultaneously in a large proportion of learners, thereby systematically increasing their shared variance. In this sense, the high correlation reflects less a lack of discriminant validity than a genuine characteristic of the sample under investigation.

Despite certain limitations (such as the AVE for FF) the global model fit indices and the theoretical consistency support the viability of the two-factor structure. Overall, the findings provide robust and relevant insights into the structure of general model understanding and constitute a solid foundation for further research.

CONCLUSION

The findings of the present study substantiate the viability of the FMCD as a theoretical foundation for describing general model understanding. Through statistical validation using CFA, the proposed two-factor model was successfully confirmed. The distinction between FG and FF is supported by the results of the study. It is thus demonstrated that general model understanding can be systematically and reliably measured along these two dimensions.

With regard to the first research question, the findings provide clear evidence for the validity of the applied test instrument in the context of general model understanding. The confirmed two-factor structure shows that the fundamental distinction between structural (visual) and functional aspects of models can be reliably captured by the item set. The instrument therefore proves to be suitable for diagnosing learners' cognitive profiles independently of domain-specific content and for classifying them within the four model types.

The empirical results of the CFA confirm the postulated two-factor structure of the FMCD also for general model conceptions. It can be clearly demonstrated that learners' mental models of scientific models can be characterized by the two dimensions FG and FF.

Another central finding of this study is that general meta-knowledge about models represents an independent cognitive demand that differs from domain-specific model competence. The prominent presence of the undeveloped type highlights that reflecting on models at a purely abstract level constitutes a greater challenge for a substantial proportion of learners than doing so within a subject-specific context. While concrete physical topics can serve as cognitive anchors that make the functional purpose of a model intuitively accessible, the general model understanding examined here requires a meta-theoretical level of abstraction that is not yet fully developed in many learners.

The analysis demonstrates that the four types (the non-developed, architectural model, dual, and functional types) can also be identified in learners' general model conceptions. Based on the levels of FG and FF, learners were successfully classified, showing that this typology is not tied to specific subject matter but instead reflects universal structural patterns.

It should be noted, however, that these findings represent primarily a snapshot and do not allow for direct causal conclusions about the underlying learning mechanisms. Nevertheless, the study provides validated scales that offer a tool for systematically evaluating the development of general model competence.

Furthermore, the study provides a plausible explanation for the high intercorrelation between the factors. This should be interpreted less as a methodological deficiency and more as an expression of the cognitive interconnections characteristic of the dual type. The dominance of this type indicates that a large proportion of learners still closely link the functionality of a model to its visual accuracy. However, this does not diminish the validity of the FF-FG framework; rather, the two-dimensional modelling makes it possible to identify this linkage as a specific learning barrier and to distinguish it from expert-level understanding (FF).

An indirect acquisition of model competence through domain-specific content alone appears insufficient. Accordingly, educators should consider that model understanding does not constitute a homogeneous construct but must be understood as multidimensional. Future research should build on these findings and, through intervention studies, examine to what extent targeted metacognitive prompts can facilitate transitions between model types. Additionally, integrating qualitative interview data could provide deeper validation of the response patterns observed here. Another promising avenue for future research would be the development of multi-tier assessment instruments based on the present questionnaire. For example, extending the instrument with additional reasoning or confidence-rating tiers could provide deeper insights into learners' underlying conceptions while reducing the influence of guessing and increasing diagnostic resolution (cf. Taşlıdere, 2016). Such an approach may complement the present short-scale design by balancing efficiency with a more differentiated assessment of model understanding.

Beyond interview-based approaches, future research could also explore mixed-methods designs by complementing the present quantitative instrument with established qualitative NOS instruments, such as the VNOS questionnaire (e.g., Abd-El-Khalick et al., 2024). In particular, open-ended questions addressing the role and purpose of scientific models could provide additional insight into learners' reasoning and help explain the cognitive profiles identified through the FG and FF dimensions. Such an approach would allow quantitative classifications to be interpreted alongside learners' own explanations of scientific models, thereby providing a richer understanding of model competence.

From an educational perspective, the validated instrument offers teachers and researchers a practical tool for diagnosing learners' understanding of scientific models before instruction. Identifying whether students primarily rely on visual similarity or functional reasoning may support the design of targeted instructional interventions that explicitly promote abstraction and model-based reasoning. Future work should investigate how such interventions influence transitions between the learner types described by the FMCD and whether similar patterns emerge across different scientific disciplines. Overall, the results once again underscore the necessity of explicitly addressing general model understanding as part of the NOS in school education.

Author contributions: HW, JV, & MSU: conceptualization, data curation, formal analysis, investigation, methodology, visualization, writing – original draft, writing – review & editing. All authors agreed with the results and conclusions.

Funding: This study was supported by the Open Access Publishing Fund of Leipzig University.

Ethical statement: This study was conducted in accordance with established ethical principles for research involving human participants. Participation was voluntary, and all participants were informed about the purpose and procedures of the study as well as the confidential handling of their data. Informed consent was obtained from all participants and, where required, from their legal guardians. Participants were informed that they could withdraw from the study at any time without disadvantage. All data were collected and analyzed in anonymized or pseudonymized form and were used exclusively for research purposes.

AI statement: Generative AI tools were used solely to improve the grammar, clarity, and phrasing of the manuscript. AI tools were not used for data collection, data analysis, interpretation of results, or the generation of scientific content. All AI-assisted revisions were reviewed and approved by the authors, who take full responsibility for the final content of the manuscript.

Declaration of interest: No conflict of interest is declared by the authors.

Data sharing statement: Data supporting the findings of this study are publicly available via Zenodo at 10.5281/zenodo.22206022.

REFERENCES

Abd-El-Khalick, F., Summers, R., Brunner, J. L., Belarmino, J., & Myers, J. (2024). Development of *VAScoR*: A rubric to qualify and score responses to the views of nature of science (*VNOS*) questionnaire. *Journal of Research in Science Teaching, 61*(7), 1641-1688. https://doi.org/10.1002/tea.21916

Ayene, M., Kriek, J., & Damtie, B. (2011). Wave-particle duality and uncertainty principle: Phenomenographic categories of description of tertiary physics students' depictions. *Physical Review Physics Education Research, 7*(2), Article 020113. https://doi.org/10.1103/PhysRevSTPER.7.020113

Bitzenbauer, P., & Fösel, A. (2025). Analysis of pre-service physics teachers' mental models of the Higgs boson using the SOLO taxonomy. *Eurasia Journal of Mathematics, Science and Technology Education, 21*(10), Article em2706. https://doi.org/10.29333/ejmste/17028

Bitzenbauer, P., & Meyn, J.-P. (2022). Toward types of students' conceptions about photons: Results of an interview study. In J. B. Marks, P. Galea, S. Gatt, & D. Sands (Eds.), *Physics teacher education. Challenges in physics education* (pp. 1875-187). Springer. https://doi.org/10.1007/978-3-031-06193-6_13

Bitzenbauer, P., & Ubben, M. S. (2025). The structure of learners' perceptions of models (not only) in quantum physics: spotlight on fidelity of gestalt and functional fidelity. *EPJ Quantum Technology, 12*, Article 16. https://doi.org/10.1140/epjqt/s40507-025-00316-7

Brown, T. A. (2015). *Confirmatory factor analysis for applied research* (2nd ed.). Guilford Press.

Chakrabarti, A., & Ghosh, J. K. (2011). AIC, BIC and recent advances in model selection. *Journal of the Indian Statistical Association, 49*(1), 1-21. https://doi.org/10.1016/B978-0-444-51862-0.50018-6

Cho, G.-H., Hwang, H., Sarstedt, M., & Ringle, C. M. (2020). Cutoff criteria for overall model fit indexes. *Journal of Marketing Analytics, 8*(4), 189-202. https://doi.org/10.1057/s41270-020-00089-1

Cohen, L., Manion, L., & Morrison, K. (2017). *Research methods in education* (8th ed.). Routledge. https://doi.org/10.4324/9781315456539

Döring, N., & Bortz, J. (2016). *Forschungsmethoden und Evaluation in den Sozial- und Humanwissenschaften* [Research methods and evaluation in the social and human sciences] (5th ed.). Springer. https://doi.org/10.1007/978-3-642-41089-5

Fornell, C., & Larcker, D. F. (1981). Evaluating structural equation models with unobservable variables and measurement error. *Journal of Marketing Research, 18*(1), 39-50. https://doi.org/10.2307/3151312

Gilbert, J. K. (2004). Models and modelling: Routes to more authentic science education. *International Journal of Science and Mathematics Education, 2*(2), 115-130. https://doi.org/10.1007/s10763-004-3186-4

Gräber, W., Nentwig, P., Koballa, T., & Evans, R. (Eds.). (2002). *Scientific literacy: Der Beitrag der Naturwissenschaften zur allgemeinen Bildung* [Scientific literacy: The contribution of the natural sciences to general education]. Leske + Budrich.

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. https://doi.org/10.1002/tea.3660280907

Kline, P. (2000). *Handbook of psychological testing* (2nd ed.). Routledge. https://doi.org/10.4324/9781315812274

Krijtenburg-Lewerissa, K., Pol, H. J., Brinkman, A., & van Joolingen, W. R. (2017). Insights into teaching quantum mechanics in secondary and lower undergraduate education. *Physical Review Physics Education Research, 13*, Article 010109. https://doi.org/10.1103/PhysRevPhysEducRes.13.010109

Lederman, N. G. (2007). Nature of science: Past, present, and future. In S. K. Abell, & N. G. Lederman (Eds.), *Handbook of research on science education* (pp. 831-879). Lawrence Erlbaum Associates.

McDonald, R. P. (1999). *Test theory: A unified treatment*. Lawrence Erlbaum Associates.

Müller, R., & Wiesner, H. (2002). Teaching quantum mechanics on an introductory level. *American Journal of Physics, 70*(3), 200-209. https://doi.org/10.1119/1.1435346

OECD. (2016). *PISA 2015 assessment and analytical framework: Science, reading, mathematics and financial literacy*. OECD Publishing. https://doi.org/10.1787/9789264255425-en

Podsakoff, P. M., MacKenzie, S. B., Lee, J.-Y., & Podsakoff, N. P. (2003). Common method biases in behavioral research: A critical review of the literature and recommended remedies. *Journal of Applied Psychology, 88*(5), 879-903. https://doi.org/10.1037/0021-9010.88.5.879

Roser, M. E., Fugelsang, J. A., Dunbar, K. N., Corballis, P. M., & Gazzaniga, M. S. (2005). Dissociating processes supporting causal perception and causal inference in the brain. *Neuropsychology, 19*(5), 591-602. https://doi.org/10.1037/0894-4105.19.5.591

Schermelleh-Engel, K., Moosbrugger, H., & Müller, H. (2003). Evaluating the fit of structural equation models: Tests of significance and descriptive goodness-of-fit measures. *Methods of Psychological Research, 8*(2), 23-74. https://doi.org/10.23668/psycharchives.12784

Schmid, J. M., Veith, J., & Bitzenbauer, P. (2026). Cognitive dimensions in students' mental models of linear light polarization. *Physical Review Physics Education Research, 22*, Article 010118. https://doi.org/10.1103/hng2-97bk

Schwarz, C. V., Reiser, B. J., Davis, E. A., Kenyon, L., Acher, 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. https://doi.org/10.1002/tea.20311

Taşlıdere, E. (2016). Development and use of a three-tier diagnostic test to assess high school students' misconceptions about the photoelectric effect. *Research in Science & Technological Education, 34*(2), 164-186. https://doi.org/10.1080/02635143.2015.1124409

Treagust, D. F., Chittleborough, M., & 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

Ubben, M. S. (2020). *Typisierung des Verständnisses mentaler Modelle mittels empirischer Datenerhebung am Beispiel der Quantenphysik* [Typification of the understanding of mental models using empirical data collection using the example of quantum physics]. Logos Verlag. https://doi.org/10.30819/5181

Ubben, M. S., & Bitzenbauer, P. (2022). Two cognitive dimensions of students' mental models in science: Fidelity of gestalt and functional fidelity. *Education Sciences, 12*(3), Article 163. https://doi.org/10.3390/educsci12030163

Ubben, M. S., & Bitzenbauer, P. (2023). Exploring the relationship between students' conceptual understanding and model thinking in quantum optics. *Frontiers in Quantum Science and Technology, 2*. https://doi.org/10.3389/frqst.2023.1207619

Ubben, M. S., & Heusler, S. (2021). Gestalt and functionality as independent dimensions of mental models in science. *Research in Science Education, 51*, 1349-1363. https://doi.org/10.1007/s11165-019-09892-y

Ubben, M. S., Hartmann, J., & Pusch, A. (2022). Mental models of students regarding black holes. *Astronomy Education Journal, 2*(1). https://doi.org/10.32374/AEJ.2022.2.1.029ra

Ubben, M. S., Veith, J. M., Merzel, A., & Bitzenbauer, P. (2023). Quantum science in a nutshell: Fostering students' functional understanding of models. *Frontiers in Education, 8*. https://doi.org/10.3389/feduc.2023.1192708

Veith, J. M., Machisi, E., Ubben, M. S., Hennig, F., & Bitzenbauer, P. (2025). Analysis of high school students' mental models of angles. *Frontiers in Education, 10*. https://doi.org/10.3389/feduc.2025.1679857

Vrieze, S. I. (2012). Model selection and psychological theory: A discussion of the differences between the Akaike information criterion (AIC) and the Bayesian information criterion (BIC). *Psychological Methods, 17*(2), 228-243. https://doi.org/10.1037/a0027127