Progress in Science Education
2025, Vol 8, No. 2, 51-69
ISSN 2405-6057
DOI.10.25321/prise.2025.1529

Research Article

How to measure Mathematical Modelling in chemistry? Using an atomistic approach for developing a new test instrument.

Benjamin Stöger[1], Claudia Nerdel[1]

Received: May 2024 / Accepted: May 2025

Structured Abstract

Background: Models and mathematical modelling play pivotal roles in understanding complex scientific phenomena. Their application ranges widely across educational and scientific disciplines, notably in natural sciences and chemistry, facilitating the translation between conceptual understanding and practical applications.

Purpose: This study aims to develop and validate a test instrument for assessing mathematical modeling competencies in the natural sciences, with a particular focus on chemistry and biology, exploring the empirical substantiation of theoretically implied subcategories and their interactions.

Sample/setting: The study involved 390 STEM students from German universities, with a final dataset comprising 309 participants. The testing occurred during the winter semester of 2022/2023.

Design and Methods: Utilizing a quantitative design, the study employed Rasch analyses based on probabilistic test theory to ensure the reliability and objectivity of the test instrument. Items were crafted to reflect various aspects of the mathematical modelling process, categorized into understanding, simplifying, mathematizing, interpreting, validating, and communicating.

Results: Findings suggest that the test instrument effectively measures mathematical modelling competencies across the specified categories. The analysis confirmed the instrument's unidimensionality, reliability, and substantive validity of the mathematical modelling constructs it aims to assess.

Conclusions: The development of this test instrument holds significant potential for educational practice and future research by providing a reliable means to assess and understand students' competencies in mathematical modelling within the natural sciences.

Keywords: *Mathematical Modeling, Natural Sciences, Chemistry Education, Test Instrument, Rasch Analysis, Assesment of competencies.*

[1]Technical University of Munich, Associate Professorship of Life Sciences Education

✉ benjamin.stoeger@tum.de

1. Scientific models in modern society

Modelling is a key tool in the natural sciences, a fact that became particularly clear during the COVID-19 pandemic in 2020. Mathematical models were used to predict the course of the pandemic and modeled forecasts were bases for political decisions. This illustrates that modelling and the use of models are essential components of scientific expertise and must, therefore, be considered fundamental to modern science education (Norris & Phillips, 2003). Müller et al. (2018) identified mathematical knowledge as a success factor in natural sciences. Mathematics therefore promotes competence in chemistry and emphasises the importance of mathematical knowledge for understanding models. However, it should be emphasised that scientific knowledge is also required to derive a valid mathematical model from an initial chemical model. Specifically, in chemistry, understanding mathematical models of molecular interactions and reaction rates requires both a solid foundation in mathematics and a deep understanding of chemical principles (Müller et al., 2018; Kimpel, 2018) findings on mathematical knowledge underline that understanding complex chemical models relies heavily on mathematical reasoning. However, mathematical concepts can present challenges for students in the context of chemistry due to the specialized skills needed to apply mathematics effectively in chemical analyses and reactions. This highlights the importance of integrating mathematical methods specifically into chemistry education.

2. Research Background

Models simplify complex phenomena and support scientific understanding, especially in chemistry where they bridge macroscopic and submicroscopic levels (Stachowiak, 1980; Johnstone, 1991). Cognitive models enable learners to internalize and apply scientific principles. As scientific complexity increases, so does the need for mathematical interpretation—emphasizing the close link between science and mathematics (Müller et al., 2018; Höhner, 1996).

This is true not only for understanding models, but also for the method of modeling itself. Modelling is a method of scientific knowledge acquisition. Additionally, model understanding is fundamental. Hodson (1992) formulated goals for scientific education which can be apadpted on modelling. He differentiated between the dimensions *learning science*, *learning about science*, and *doing science*. In addition to the first two more declarative dimensions, the third dimension describes methods of knowledge acquisition such as experimenting or *modeling*. For Modelling as part of the inquiry process - which involves systematically exploring questions, gathering and analyzing data, and building understanding through evidence- many empircal research has been made. All of them focus on an overall model competence and modelling competence (Leisner, 2005; Mayer, 2007; Meisert, 2008; Upmeier zu Belzen & Krüger, 2010; Stäudel, 2012). It is demonstrated that, alongside general and subject-specific understanding of models and modeling competence, the skill to apply modeling methods is essential for effective *modeling*. With regard to the mentioned significance of mathematics in scientific models, it becomes evident that mathematics plays an crucial role in fostering modeling skills.

2.1 Mathematical Modelling

In mathematics, *Mathematical Modeling* and *Mathematization* are fundamental skills that enable the translation of real-world problems into mathematical language. *Mathematization* involves identifying and representing relationships and patterns in mathematical terms, such as symbols, equations, or diagrams (Zech, 1998). In contrast, *Mathematical Modeling* is a more comprehensive process, encompassing not only mathematization but also applying mathematical tools to analyze, interpret, and solve problems (Greefrath, Kaiser, Blum, & Borromeo Ferri, 2013). *Mathematical Modeling* is, therefore, understood as transforming concrete, application-driven problems into structured, solvable mathematical tasks (Eck, Garcke, & Knabner, 2017). This complexity requires specific sub-competencies within mathematical modeling, such as Understanding, Simplifying, Mathematizing, Interpreting, Validating, and Communicating (Greefrath, Kaiser, Blum, & Borromeo Ferri, 2013).

The process of mathematical modeling in mathematics is often described by structured frameworks or cycles that break down the steps to create and work with models effectively. One example is the modeling cycle proposed by Blum and Leiss (2005), designed to identify and describe aspects of students' process of modeling (See fig. 1). This cycle starts with understanding the problem in a real-world context and creating a *Situational Model* (or mental model) based on the learner's interpretation of the problem. The modeler simplifies this situational model to highlight relevant information, ignoring less significant aspects. This results in a *Real Model* of the problem, which can then be translated into a *Mathematical Model* through mathematization, allowing mathematical tools to solve or manipulate the model.

After obtaining mathematical results, the modeler interprets these within the original real-world context, ensuring they correspond to practical, relevant outcomes. Finally, the results are validated against the initial mental model, confirming the accuracy and applicability of the solution. If discrepancies arise, the modeler may need to refine the situational model and repeat the cycle until a satisfactory outcome is achieved (Blum & Leiss, 2005). This cycle offers a systematic approach, emphasizing the distinction between mathematical and real-world contexts and fostering a deep understanding of mathematical modeling as an iterative, reflective process.

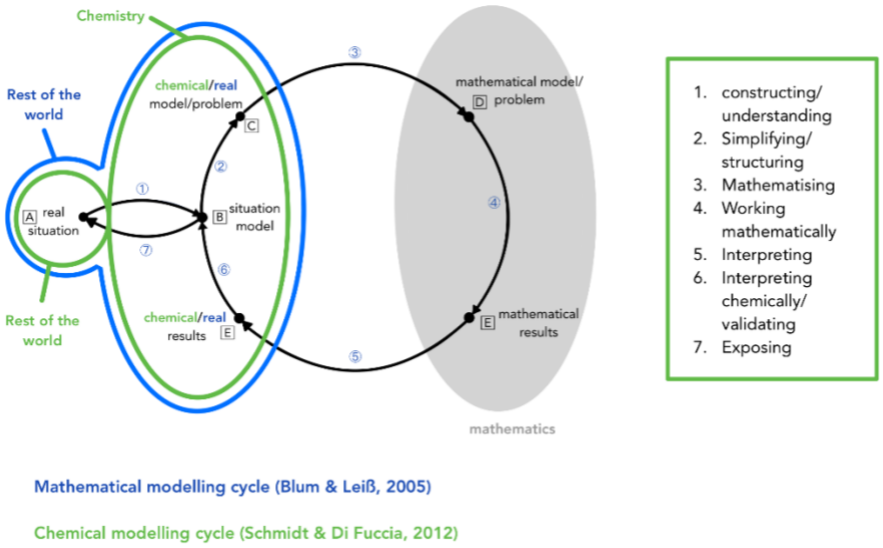

Fig. 1. Mathematical modeling cycles according to Blum and Leiss (2005) and Schmidt and Di Fuccia (2013).

2.2 Mathematical Modelling in Chemistry

For chemistry, *Mathematical Modeling* cycles were, adapted from mathematical modeling cycles to accommodate the specific needs of chemical contents. Schmidt and Di Fuccia (2013), therefore, adopted it for application in chemistry, positioning the Chemical Model as a bridge between the real world (referred to as the "Rest of the World") and mathematical models (See fig. 1). Rest of the World describes the macroscopic real world problem. Unlike in mathematics, where the situational model directly informs the mathematical model, chemistry requires an intermediate step to create a chemical model that reflects the underlying submicroscopic chemical processes (Goldhausen, 2015). This reflects the need for chemical explanations—often involving molecular or atomic-level insights—that serve as an essential bridge between reality and mathematics (Johnstone, 1991). The process in chemistry begins with constructing a situational model and then simplifying and structuring it to develop a *Chemical Model*. This model often takes forms specific to chemistry, such as reaction equations or molecular formulas, which provide a framework to represent the problem within chemical principles (Goldhausen, 2015; Kimpel, 2018). After developing a chemical model, mathematics can be applied, treating it as a quantification tool rather than the core focus of analysis. The chemical model is therefor translated into a *Mathematical Model*, maintaining fidelity to the chemical context. This allows the problem to be analyzed mathematically, producing results that can be translated back to a chemical interpretation.

The modeling cycle in chemistry proceeds with interpreting the mathematical results within the chemical context to ensure they align with the initial situational model and chemical understanding (Goldhausen, 2015). If the outcome is coherent from a chemical perspective, it can be applied back to the real-world problem. If discrepancies appear, the modeler revisits the situational model, re-evaluating the chemical and mathematical processes involved, and iterates as necessary to reach a valid solution (Schmidt & Di Fuccia, 2013). This iterative process highlights the distinct role of mathematics as a supportive tool within chemistry, where it serves to quantify and analyze data within chemically defined frameworks.

2.3 Differences and similarities

Although mathematical modeling in both mathematics and chemistry follows an iterative, cyclical process, key differences reflect the specific goals and frameworks within each discipline. In both cases, modeling begins by understanding the real-world situation and creating an initial situational model. However, in mathematics, the situational model is quickly simplified and mathematized, leading directly into mathematical manipulation. The mathematical results are then interpreted in real-world terms, with any necessary refinements to the situational model based on the fit of these results to the original problem context. This cycle, as shown in the Blum and Leiss (2005) model, focuses heavily on translating real-world problems directly into mathematical forms and interpreting results within this mathematical framework.

In chemistry, however, mathematical modeling is more layered, requiring a chemical intermediary step between the real-world situational model and the mathematical analysis. This difference is significant, as chemical modeling involves submicroscopic explanations (such as molecular interactions) that are not directly represented through mathematics but are essential to the model's relevance and accuracy. Consequently, in chemistry, mathematics serves primarily as a quantification tool applied only after a robust chemical model has been established. This additional layer means that chemistry modeling cycles, like those described by Schmidt and Di Fuccia (2013), integrate a translation back into the chemical context to ensure the results align chemically before they are applied to the real-world problem.

In summary, while mathematical modeling in both fields involves structured cycles of understanding, simplifying, translating, and validating, mathematics uses a direct path from real-world context to mathematical representation. In chemistry, the cycle is more complex, requiring an intermediate chemical model that integrates both chemical knowledge and mathematical tools. This distinction highlights mathematics' role as both a conceptual foundation in mathematics itself and a practical tool for quantification in chemistry, thus underscoring how modeling processes are tailored to the disciplinary demands of each field.

2.4 Measuring mathematical modelling

While both chemistry and mathematics employ modeling cycles to translate real-world phenomena into a usable form, the approach in mathematics education research emphasizes evaluating and developing specific modeling skills. This structured approach in mathematics education has led to the development of test instruments aimed at assessing learners' proficiency in various phases of the modeling cycle. For instance, Haines, Crouch, and Davis (2001) initiated the investigation of learners' modeling abilities with targeted tasks that isolate key components of mathematical modeling, such as problem formulation, variable assignment, and model selection. These studies reflect a foundational shift towards identifying critical sub-competencies within the modeling cycle.

An important theoretical framework for these efforts stems from holistic and atomistic approaches to mathematical modeling. Cevikbas et al. (2022) emphasize the structured nature of these approaches, noting their capacity to capture the multifaceted competencies involved in mathematical modeling. Hidayat et al. (2022) confirm the overrepresentation of holistic methods in the literature and highlight the need for greater emphasis on atomistic approaches. They argue that atomistic methods hold significant potential for advancing test instruments through their diagnostic precision and standardized applicability. The holistic approach, as highlighted by both earlier and recent studies (e.g., Brand, 2014; Czocher et al., 2021), focuses on evaluating the entire modeling process, including the integration of real-world contexts with mathematical abstraction and subsequent interpretation. This perspective has proven particularly effective for assessing how learners navigate the interconnected phases of the modeling cycle and apply mathematical solutions to realistic scenarios. Moreover, studies discussed in Czocher et al. (2021) underline the potential of holistic methodologies to reinforce connections between theoretical understanding and practical applications, especially in interdisciplinary fields like engineering and natural sciences.

For this purpose, various tasks were designed within the modelling cycle, focusing on different areas of mathematical modeling. The items were grouped into the following categories: *formulating the problem; assigning variables, parameters, constants; formulating mathematical formulas; choosing a model*. Houston and Neill (2003) added two additional item categories to these four, one of which included the use of graphical representations and the other focused on the return to the real problem situation. This research laid the foundation for many further studies, which dealt with different aspects and possibilities of capturing mathematical modeling (Haines et al., 2001; Lingefjärd & Holmquist, 2005; Kaiser & Schwarz, 2006; Izard, 2007; Zöttl, Ufer, & Reiss, 2011; Grünewald, 2013; Borromeo Ferri et al., 2013; Fu & Xie, 2013; Brand, 2014). For example, atomistic items were compared with holistic items (Zöttl et al., 2011; Brand, 2014). This perspective, championed by Cevikbas et al. (2022) and supported by Hidayat et al. (2022), facilitates a more precise diagnosis of learners' strengths and weaknesses in distinct modeling phases. Their systematic review identifies a gap in the overrepresentation of holistic methods and argues for the development of standardized atomistic instruments to ensure diagnostic clarity.

For the atomistic approach, the steps within the corresponding modeling cycles were operationalized and items were developed for them. In the holistic approach, the entire modeling cycle was the focus. Therefor items were developed based on sub-competencies of mathematical modeling (Maaß, 2004; Kaiser, 2007; Kaiser & Schwarz, 2006, 2010; Zöttl et al., 2011; Brand, 2014; Hankeln, Adamek, & Greefrath, 2019).

In their atomistic approaches, studies differentiate between various sub-aspects of modeling cycles (Figure 2) using items that focus on the transition from reality to mathematics and from mathematics to reality (1 & 3 in Figure 2) (Houston & Neill, 2003; Zöttl et al., 2011; Hankeln et al., 2019). Additionally, they include items focusing on mathematical operations (2) (Brand, 2014). The items in the first group (Fig. 2, Step 1) focus on competencies necessary for understanding a real problem and for establishing a real model, as well as competencies for generating a mathematical model from a real model (Zöttl et al., 2011; Hankeln et al., 2019). The second group (Fig. 2, Step 2) contains competencies for solving mathematical problems within a mathematical model(Brand, 2014; Hankeln et al., 2019). Lastly, the third group (Fig. 2, Step 3) of items focuses on competencies for interpreting mathematical results within a real model, as well as competencies for questioning solutions and possibly re-initiating the modeling process (Zöttl et al., 2011; Brand, 2014). Hidayat et al. (2022) stress that operationalizing these steps not only supports theoretical rigor but also provides a practical framework for constructing standardized test instruments in the natural sciences. This methodological clarity, as supported by probabilistic test theory (Zöttl et al., 2011; Brand, 2014), enables the construction of multidimensional models that effectively capture diverse competencies in mathematical modeling.

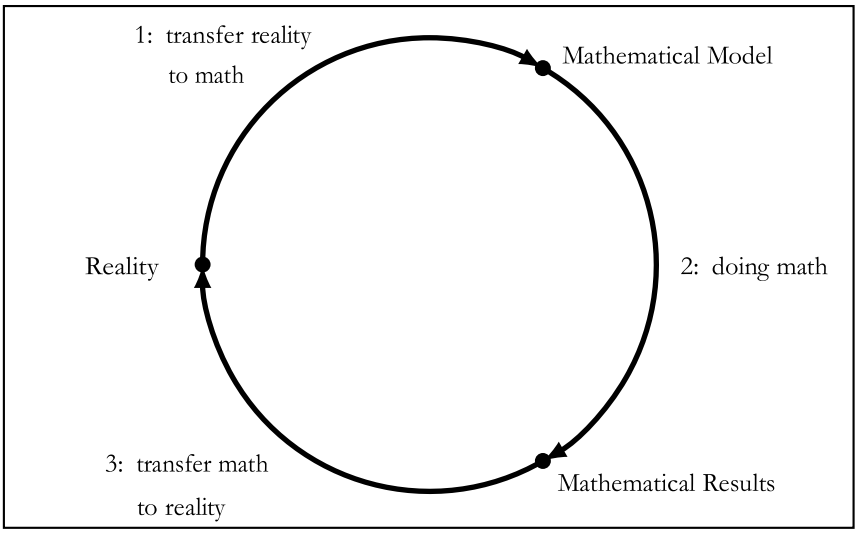

Fig. 2. Idealized schematic structure of the categories and steps of mathematical modeling that served as the basis for item construction.

Fundamentally, these steps always differ in the way that they consider the transition between the levels of Reality and Mathematics (Zöttl et al., 2011; Brand, 2014). Additionally, a few studies supplemented a category of overall modeling. This integrated a holistic task type in contrast to the atomistic approach. Thus, this additional category targets the ability to solve complete modeling tasks, as well as the metacognitive aspects of the solution process of modeling tasks (Haines et al., 2001; Zöttl et al., 2011; Brand, 2014). Almost all studies validated their instruments using probabilistic test theory and concluded that atomistic items are also suitable for tests to capture mathematical modeling (Zöttl et al., 2011).

By operationalizing transitions between reality and mathematics (Steps 1 and 3 in Fig. 2) and mathematical problem-solving (Step 2), atomistic items provide a granular understanding of learners' abilities at each stage.

Using probabilistic test theory, has also shown that a multidimensional model, where each step within the modeling process (see fig. 2) loads onto its own dimension, best represents the captured data (Brand, 2014).

These findings from mathematics education research, combined with the modeling cycle for chemistry (Schmidt & Di Fuccia, 2013; Goldhausen, 2015), provide the theoretical foundation for the development of a test instrument for mathematical modeling in chemistry. The integration of both holistic and atomistic perspectives, as suggested by Cevikbas et al. (2022) and Hidayat et al. (2022), underscores the potential to create a balanced framework that leverages the strengths of both approaches for assessing modeling competencies in chemistry. Hidayat et al. (2022) particularly emphasize the importance of balancing these perspectives to address the diverse requirements of interdisciplinary fields like natural sciences, while also advocating for the development of standardized, validated test procedures.

Building on this foundation, it becomes crucial to explore whether such a balanced framework can be translated into a structural model for mathematical modeling specific to natural sciences, such as chemistry. This raises theoretical and methodological questions about how to operationalize and empirically validate the subcategories of mathematical modeling within these disciplines.

On a theoretical and methodological level, the question arises whether a similar structural model for mathematical modeling in the natural sciences (chemistry/biology) can be developed. Accordingly, it is necessary to investigate whether theoretically implied subcategories of mathematical modeling are empirically proofable and how they interact with each other. From this question, the following hypotheses were developed: (H) Mathematical modeling in the natural sciences can be reliably measured by operationalizing the individual steps within the mathematical modeling cycle.

3. Methods

3.1 Preliminary Considerations

Primarily, the developed instrument should target students at a university level. To develop an instrument with a broad disciplinary range, different subfields within the discipline of chemistry need to be considered. These subfields (Chemical Reactions, Stoichiometry, Physical Chemistry, and Biochemistry) were examined to determine the extent to which mathematizations and mathematical models are present and utilized. While content from both school and university curricula was reviewed to ensure comprehensive coverage, the primary focus of this instrument is on university-level. This examination included content from both school and university curricula. Accordingly, various subject-specific contents were identified in which mathematizations or mathematical models could be found (Table 1).

It was found that in almost all areas of chemistry, mathematization or mathematical models are utilized. This includes fundamental aspects of chemical reactions such as the Law of Conservation of Mass, the Law of Mass Action, and the determination of amounts of substances and molar concentrations. However, more complex topics also exist that are based on mathematical models. Examples include various atomic models, thermodynamic considerations of chemical reactions, and studies on reaction rates, extending to complex models such as the Michaelis-Menten kinetics or the Schrödinger Equation. These topics were specifically added because of their relevance to university students in chemistry, ensuring that the instrument is suited to assessing the skills expected at this educational level. Subsequently, mathematical contents/skills/competencies associated with these mathematizations, and mathematical models were identified. Various mathematical contents of differing complexity were recognized and compiled (Table 2). Therefore, we were able to delineate both relevant chemical contents and the corresponding mathematical contents necessary for a comprehensive instrument.

Tab. 1. Chemical Subfields and Specific Topics for Mathematization and Mathematical Models (not exhaustive)

Chemical Subfield	Specific Topics for Mathematization/Mathematical Models
Chemical Reactions	Reaction equations, Conservation of Mass
Chemical Equilibrium	Law of Mass Action, Solubility Equilibrium, Le Chatelier's Principle
Scientific Methodology	Analysis of experiments, Mixing Cross, Chemical Analytics
Atomic/Molecular Structure	VSEPR, Symmetry, Orbitals, Bohr's Atomic Model
Stoichiometry	Determination of Substance Amounts/Concentrations, Empirical Formulas, Molar Masses
Organic Chemistry	Hückel's Rule, Phenomenon of Optical Activity
Biochemistry	Michaelis-Menten Kinetics, Enzyme Inhibition
Acid-Base Chemistry	$pH/pK_a/pK_B$ Value, Titrations, Buffer Equation
Thermodynamics	Enthalpy, Entropy, Free Enthalpy, Activation Energy
Electrochemistry	Galvanic Cell, Galvanic Series, Nernst Equation, Faraday's Laws
Reaction Kinetics	Reaction Rate, Reaction Order
Color	Absorption, Emission, Beer–Lambert law
Redox Reactions	Oxidation state

Therefore, we were able to delineate both relevant chemical contents and the corresponding mathematical contents necessary for a comprehensive instrument.

Tab. 2. Mathematical Contents Relevant to the Chemical Context (not exhaustive)

Subject Area	Contents
Calculus	Derivatives, Differential Calculus, Integral Calculus, Finding Indefinite Integrals, Differential Equations
Algebra	Solving Linear, Quadratic, Polynomial, Rational, Exponential, Logarithmic Equations, Reformulating Terms, Evaluating (Function) Terms, equivalent transformation, Concept of Functions, Direct/Inverse Proportionality, Solving systems of linear equations, Graphical Evaluation of Functions, Percentage Calculations, Basic Arithmetic Operations, Symmetric Groups
Geometry	Area Calculation, Calculating Surface Areas and Volumes, Identifying Types of Symmetry, Symmetry Operations, Area Calculation Using Definite Integrals
Statistics	Relative Frequency, Conditional Probability, regression analysis, measurement errors

Theoretical Conception

The development of the test instrument is based on the modelling cycle in chemistry as proposed by Schmidt and Di Fuccia (2013) and Goldhausen (2015). This framework builds on the general modelling cycle by Blum and Leiss (2005) and adapts it to the specific demands of chemical modelling by introducing an intermediate chemical model between real-world problems and mathematical representation. The modelling process is structured around three domains: the real world (I), chemistry (II), and mathematics (III), with transitions between them serving as the basis for six categories (A1–A6).

- A1 (Rest of the World → Chemistry) includes both constructing and simplifying a situational model into a chemical model. Items address the identification of relevant features, the choice of appropriate model conceptions, and the structuring of content on a conceptual level (cf. Borromeo Ferri, 2004; Schmidt & Di Fuccia, 2013).
- A2 (Chemistry → Mathematics) refers to mathematization. Tasks focus on recognizing mathematical structures, developing or selecting formulas, and translating relationships into mathematical representations (cf. Kaiser & Schwarz, 2006, 2010; Brand, 2014).
- A3 (Working mathematically) assesses the application of mathematical methods such as equation solving, calculus, and function analysis. Items in this category were designed across a range of difficulty levels and mathematical subdomains (cf. Hankeln, Adamek & Greefrath, 2019).

- A4 (Mathematics → Chemistry) focuses on interpreting mathematical results within the chemical model. Items assess the assignment of units or variables, translation into chemical language, and the selection of relevant results (cf. Goldhausen, 2015).
- A5 (Chemistry → Rest of the World) addresses the transition back to the macroscopic level. Items target reflection, validation of results, contextual interpretation, and communication of findings (cf. Schmidt & Di Fuccia, 2013).
- A6 (Overall Modelling) integrates all modelling phases and emphasizes meta-level skills such as process planning, reflection, and justification (cf. Zöttl, Ufer & Reiss, 2011; Brand, 2014).

Each category was defined to reflect distinct steps of the modelling cycle and to align with theoretical frameworks from both mathematics and chemistry education (Borromeo Ferri, 2004, 2006; Zöttl et al., 2011; Kaiser & Schwarz, 2006, 2010; Brand, 2014; Hidayat et al., 2022; Cevikbas et al., 2022). The assignment of tasks to categories followed specific item construction criteria as described by Rost (2004), ensuring content validity, clarity, and diagnostic relevance.

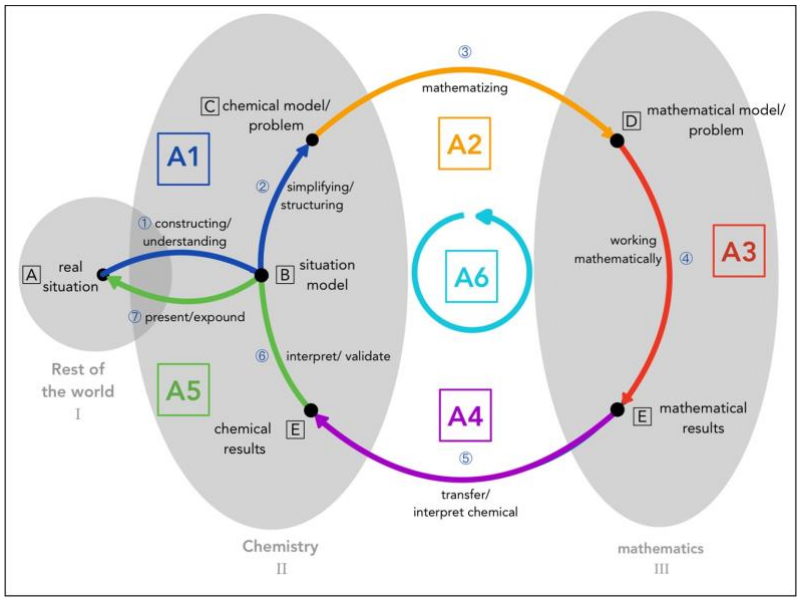

Fig. 3. Theoretical Conception of the Individual Item Categories A1 to A6 based on Schmidt & Di Fuccia (2013).

3.2 Item and Test Booklet Construction

Item development was guided by the theoretical framework and tailored to each category's modelling focus. To ensure construct validity, experts from chemistry and mathematics education reviewed item content and structure. Their feedback was systematically integrated into item refinement. The same chemistry topics were used across all categories, but each task was adapted to align with the specific modelling step, minimizing content bias between categories. For categories A1, A2, A4, and A6, 12 items each were constructed. With the help of this amount of items per category, the content described in section 3.1 and the task types described in 3.2 could all be included. For category A6, a total of six items were developed, and for category A3, a total of 12 groups of 3 items each (36 in total) were created. For each of the 12 groups of items in Category A3, three different difficulties of items were generated. That's the reason why there are 36 items in category A3. Overall there were 90 contextual items generated ($N_c = 90$). All these items were developed following the principles of item construction (Rost, 2004). A single choice format was chosen. The items for categories A1, A2, A4, A5, and A6 had an identical structure. Each item began with a task description in text form (supported by additional representations, such as images or diagrams, if necessary). This was followed by five different response options: one correct answer, two incorrect answers, and two plausible answers. The plausible answers included common misconceptions or errors frequently made by learners (Barke, 2006; Göhner & Krell, 2021). Both text-based and image-based response formats were developed. The following images show one sample item each from categories A1 to A5. No sample item has been added for A6 due to the length of the item.

Unlike the other categories, items in category A3 were formulated as dichotomous items. This decision was made in order to also use category A3 as a mathematical knowledge test. This enables a clear classification of mathematical skills. For category A3, 12 mathematical subject areas were selected (e.g., Symmetry, Area & Volume Calculation, Functions, Differential Calculus). For each of these areas, three items of varying difficulty (increasing abstraction) were developed. Each item in this category has five answer options, of which only one is correct. In addition to the subject-specific items, items were generated to capture personal variables such as age, gender or course of study ($N_p = 15$). Due to the large number of contextual items ($N_c = 90$), twelve test booklets were designed (Frey, Hartig, & Rupp, 2009). Each test booklet included 23 items in total, distributed according to the following scheme: three items each from categories A1, A2, A4, and A5, as well as three topics (nine items) from category A3, and two items from category A6.

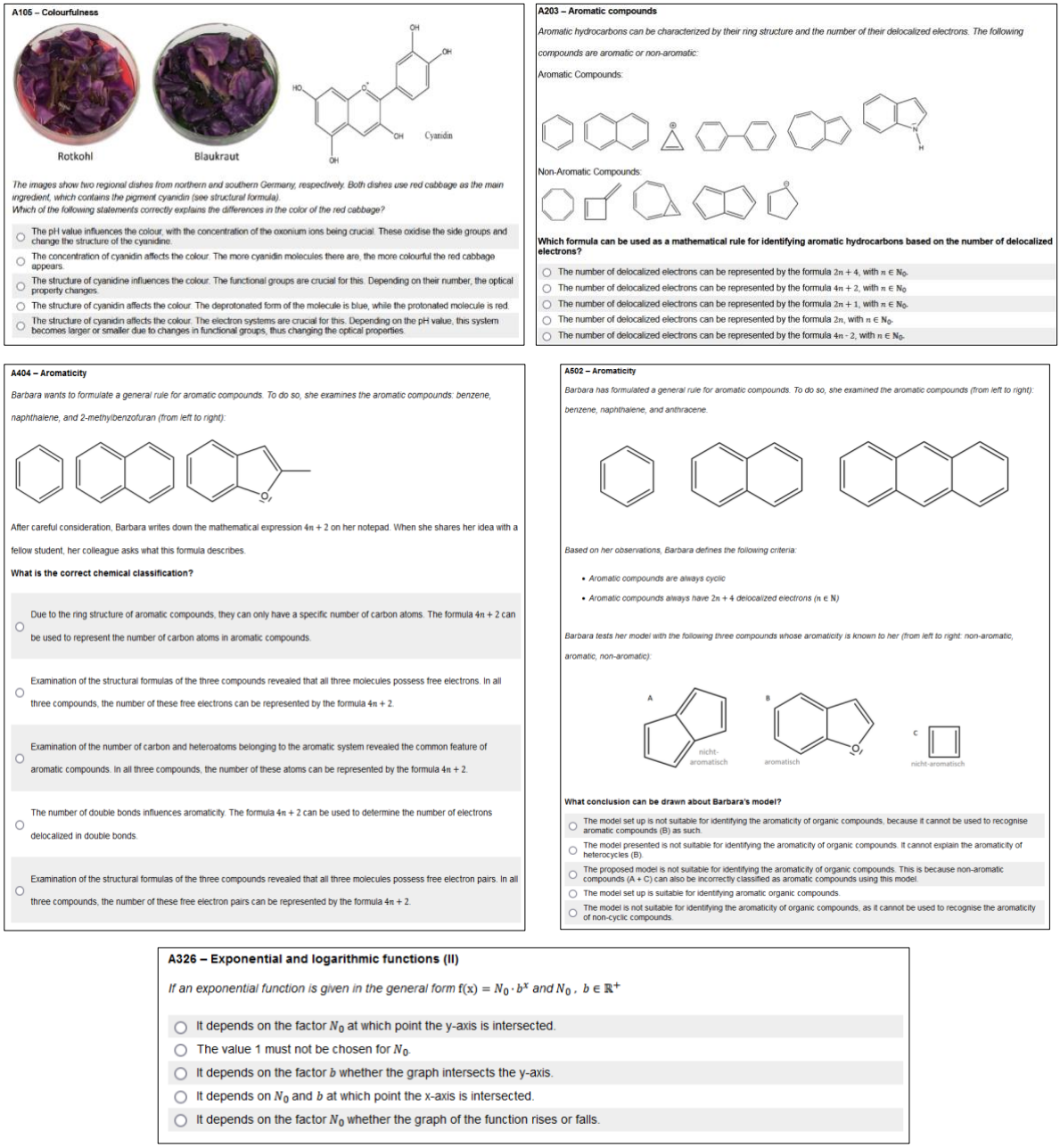

Fig. 4. Example items for categories A1 to A5.

3.3 Methodological Approach

The test booklets were distributed as an online questionnaire (SociSurvey), each containing 23 items with no time limit. Data collection took place during the winter semester 2022/2023. A total of N=390 students from German universities in the fields of natural sciences, mathematics, or engineering participated. The final dataset included 309 participants; those who answered less than 20% of the items were excluded.

Participants were on average 22.5 years old (SD = 5.45; Median = 21). Gender distribution was 47.9% male, 48.2% female, 1.3% diverse, and 2.6% undisclosed. On average, they were in their fourth semester (M = 3.70; SD = 3.12). All were enrolled in MINT degree programmes. The subject distribution was: Chemistry (33.98%, n=105), Biology/Biochemistry/Biotechnology (20.38%, n=63), Education (17.47%, n=54), Physics (11.32%, n=35), Mathematics (7.11%, n=22), Engineering (2.58%, n=8), with other or unspecified fields comprising 7.11% (n=22).

Using this dataset, methods of probabilistic test theory, specifically Rasch analyses, were conducted. The Rasch model allows for the quantification of latent traits such as abilities or attitudes on a onedimensional scale by modeling the probability of item response in relation to person ability and item difficulty. Using the Rasch equation (1, where $P(X_{ni} = 1)$ is the probability that person n correctly answers item i, θ_n represents the ability of person n, and β_i the difficulty of item i), a response probability for individual items and persons can be derived.

$$P(X_{ni} = 1) = \frac{\exp(\theta_n - \beta_i)}{1 + \exp(\theta_n - \beta_i)}$$

For the application of the Rasch model, the test items were constructed to fit the assumptions of unidimensionality and local independence (Chapter 3.3). Unidimensionality implies that the items measure a single latent trait (here: Mathematical Modeling). Local independence means that the responses to the items are independent of each other.

Additionally, Rasch analyses follow the principle of objectivity, ensuring that the ability of individuals is independent of the specific items they answered, and vice versa. This allows independent evaluation and comparison of items and persons.

The calculation of person abilities and item difficulties from the raw data also linearizes the data. While the raw data of an ordinal scale only suggest linearity through their coding, the calculated person abilities and item difficulties are on a linear scale. Moreover, missing data in probabilistic test theory is not as problematic as it is in Classical Test Theory. The Rasch model computes missing values by using the Rasch equation. This allows it to compute values based on item difficulties and person abilities. In comparison to classical test theory, where missing values were generated based on the mean of the other participants answers. However, these advantages also impose strict prerequisites for conducting a Rasch analysis. Unlike Classical Test Theory, whose assumptions are taken as untested axioms, Probabilistic Test Theory postulates test or measurement models that are inferentially tested with the data collected from a test. Only if the data of all items conform to the measurement model, the test values can be accepted as true measurements of the latent construct and used for diagnostic or research purposes.

The Rasch analysis was conducted using the software *Winsteps* (Version 5.3.4.0), where the quality of the instrument is assessed using various aspects from this analysis.

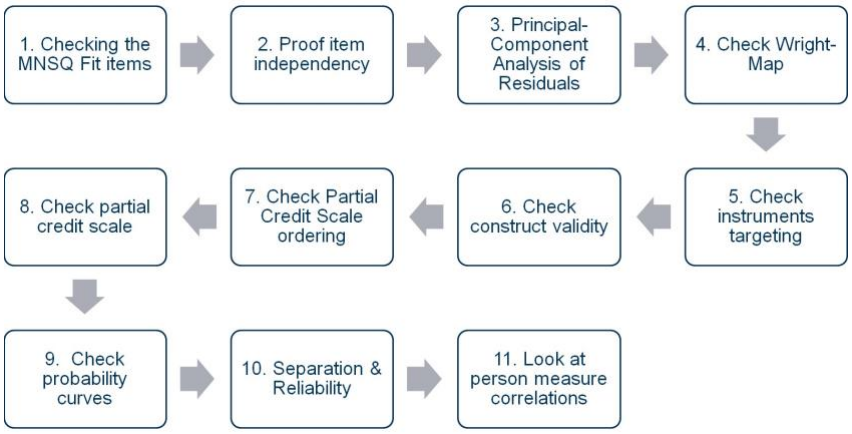

Fig. 5. Aspects Examined in the Context of Rasch Analysis and Test Development.

This methodological approach provides a robust foundation for evaluating the measurement quality of the instrument. In total, a Rasch analysis was conducted for each individual category (A1 to A6) to justify the unidimensionality and independence of each subcategory. Additionally, a Rasch analysis for the entire instrument with all 90 items was carried out to examine the unidimensionality of the instrument. Subsequently, various aspects of the Rasch analyses were examined to assess the quality. The specific approach, along with all aspects, is depicted in Figure 4.

The most critical points in this context were the verification of unidimensionality (1,3), local statistical independence (2), reliability (10), the quality of the response scale (7,8), and item targeting (4,5).

Since unidimensionality is a central assumption of the Rasch model, it is verified based on the model fit of individual items. Item fit statistics are calculated by summing and squaring the deviations of observed values from expected values. The mean value of these deviations (MNSQ value) reflects the item fit statistic. MNSQ values greater than 1.0 suggest a deviation from unidimensionality, while values less than 1.0 indicate redundancy and potential violation of local stochastic independence. Fit statistics between 0.5 and 1.5 are suitable for a multi-level response format (Boone, Staver, & Yale, 2013; Malec et al., 2007), and for samples with $N < 1000$, an MNSQ value in the range of 0.5 to 2.0 is considered acceptable (Wolfe & Smith Jr, 2007). Items that fall outside this range need to be re-examined for content-related criteria (wording, misleading question formulation, the goal of the task does not match the latent construct) and may need to be revised or excluded. Winsteps differentiates between two types of fit statistics: Infit and Outfit. Both are χ^2 tests that provide information about the relationship between data and the Rasch model. Infit and Outfit consider different aspects. Infit checks whether persons with an ability similar to the difficulty of the item answer it as expected, while Outfit checks this for persons whose ability is significantly higher or lower than the difficulty of the item. Hence, Outfit is the more important measure as it weights extreme outliers more heavily (e.g., persons with high ability answering a simple item incorrectly) (Boone et al., 2013). This also helps to strengthen construct validity. Ensuring that all items fit well within this range helps confirm that they consistently measure the intended unidimensional trait of mathematical modeling in chemistry.

A Principal Component Analysis of Residuals (PCAR) is used to further investigate unidimensionality. This involves subjecting the residuals, the differences between observed data and the values predicted by the model, to another factor analysis to determine if systematic variances in the residuals exist that could indicate additional dimensions (like a different theoretical construct being involved). However, the PCAR is not interpreted in the same way as a classical factor analysis. It merely has an indicative character and can suggest an additional dimension. In PCAR, the hypothesis being refuted is that the residuals are random noise, by finding the component that explains the greatest amount of variance in the residuals. This component is the "first contrast" (or the first PCA component in the correlation matrix of residuals). If the eigenvalue of the first contrast is small (< 2.0), then the first contrast is below the noise level and the hypothesis of random noise is not generally refuted (Linacre & Wright, 2000; Boone et al., 2013).

Local statistical independence, the second central requirement for Rasch analyses, is verified by examining the pairwise correlation of the residuals of the items (2). If two items show a high correlation of their residuals, this could indicate that they are not locally independent, which would violate the assumptions of the Rasch model and could be a problem for the accuracy and validity of the scale. A high correlation is spoken of when two items share more than half of their variance (Linacre & Wright, 2000). As the correlation squared equals the shared variance of two items, the threshold for the correlation of residuals is set at 0.7. If the correlation is greater than 0.7, it must be assumed that there is local dependence and thus a violation of the assumptions of the Rasch model (Linacre & Wright, 2000; Boone et al., 2013). There are two possible reasons for high correlations of residuals: either the duplication of a feature or the inclusion of a shared dimension. Only if local statistical independence is present a Rasch analysis can be conducted.

The reliability of the instrument and its subcategories was checked using two statistical criteria: the separation index and (separation) reliability. The separation index is the ratio between variance and error variance (Linacre & Wright, 2000). The separation reliability is comparable to reliabilities from classical test theory, like Cronbach's α (Boone et al., 2013). Item reliability is independent of test length and largely unaffected by model fit. Item separation is used to check the item hierarchy. A low item separation (< 2.5; reliability < 0.8) suggests a sample size that is too small (Boone et al., 2013; Tennant & Conaghan, 2007). Reliability is a measure of the reproducibility of the relative measurement position. This means that with high reliability, it is likely that items estimated with high measurement values actually have higher measurement values than items with low measurement values. The reliability of the estimation increases with increasing sample size and/or with low measurement errors. Accordingly, item reliability depends on two factors. On the one hand, it depends on the variance of item difficulty: the higher the variance, the higher the item reliability. On the other hand, it also depends on the sample size: the larger the sample, the higher the item reliability (Boone et al., 2013; Boone & Staver, 2020).

The quality of the Partial-Credit Scale (7,8) was evaluated using Davis and Boone's (2021) criteria. A functional scale must meet the following:

1. Each category must have ≥10 responses.
2. Responses should cluster centrally and taper at extremes.
3. Average person ability should increase with each category.
4. Outfit MNSQ < 2.0 (ideal: 0.5–1.5).
5. Andrich thresholds must increase monotonically.
6. There must be reciprocal coherence (>40%) between category usage and person ability.
7. Category distances > 1.4 logits.
8. Category distances < 5.0 logits.

Targeting (4,5) helps to determine how appropriate the items are for the sample. For this purpose, the Wright Map is examined, which allows person abilities and item difficulties to be displayed on the same scale. The Targeting is calculated by looking at the absolute value of the difference between mean item difficulty und mean person ability. On the Wright Map, difficult and easy items can be identified based on their position on the scale (difficult items at the top, easy items at the bottom). It also allows the identification of persons with high or low abilities (high ability at the top, low ability at the bottom). The comparison of items and persons then makes it possible to make further statements about which persons would have answered which items correctly. The Wright Map provides essential support for evaluating construct validity by producing an ordered hierarchy of item difficulties. Construct validity is reinforced if items with theoretically higher difficulty levels align with their placement on the Wright Map. This alignment indicates that the instrument consistently measures the construct of mathematical modeling across items and individuals, in line with theoretical expectations. By visually confirming that the most difficult items are answered only by the most capable individuals (and vice versa), the Wright Map validates that the instrument captures variations in mathematical modeling ability as intended. The Wright Map also directly compares the mean values of person abilities and item difficulties. This comparison determines how well the sample matches the instrument. For this purpose, the absolute difference between the mean values, the so-called targeting, is considered. The smaller the difference between the mean values, the more suitable the instrument is for this sample. In ideal targeting, the average person ability and the average item difficulty are identical. For still acceptable targeting, the difference should be less than one (Linacre & Wright, 2000; Boone et al., 2013; Boone & Staver, 2020). In addition, Wright maps are a good tool for discussing the content of items. The item hierarchy provided by Rasch analyses can now be used as a basis for theoretically investigating aspects of content. In this way, content-related similarities between difficult items and their differences to simple items can be worked out very well. This also enables a more in-depth discussion of learners' abilities and skills as well as potential problems or competence levels (Boone et al., 2013; Boone & Staver, 2020).

4. Results

4.1 Statistical Findings

The participant's answers were coded in a partial-credit scale. Correct answers were coded with 2, plausible answers with 1, and incorrect answers with 0. The unidimensionality was individually examined for each category (A1-A6). For this purpose, the MNSQ Outfit of individual items within their categories was observed. Thus, items that had an MNSQ Outfit outside the acceptable interval ([0.5;1.5]) were identified. Category A4 is exemplarily shown in Table 3. There, an item with misfit can be identified. This item, A407, was therefore excluded from further analysis. This item is also a very difficult item with an item difficulty of 1.57. The item deals with the stoichiometry of chemical reaction equations. Categories A1, A2, and A6 had no items with misfit (Outfit MNSQ), in category A3 six items, and in category A5 one item were identified. By removing these items, the unidimensionality of the categories could be ensured. Moreover, this approach improved the quality of the scale. When re-examining the categories after removing these items, no further items with misfit could be identified. Due to text length, there are fewer responses in categories appearing at the end of the booklets (A4 to A6) than in categories at the beginning of the booklets (A1 to A3).

Tab. 3. Winsteps Table 3.1 - Item entry (for N=309).
Total Count is the amount of answers given to each item, JMLE measure is the item's difficulty.

Total Score	Total Count	JMLE Measure	Model S.E.	Infit		Outfit		Name
				MNSQ	ZSTD	MNSQ	ZSTD	
80	55	-.88	.23	.95	-.91	.81	-.76	A401
81	56	-.82	.23	.93	-.35	.94	-.16	A402
49	43	-.37	.24	.94	-.26	.89	-.46	A403
33	44	1.14	.23	1.33	1.75	1.35	1.55	A404
54	51	.52	.23	.75	-1.44	.77	-1.18	A405
56	51	.05	.22	1.05	.36	1.03	.22	A406
24	50	1.57	0.25	1.38	1.70	2.10	2.82	A407
60	47	-.15	0.23	0.98	-.04	1.02	.16	A408
53	42	-.33	0.25	0.87	-.63	.80	-.81	A409
51	49	.33	0.22	0.97	-.13	.95	-.22	A410
63	45	-.63	0.25	0.84	-.78	.87	-.48	A411
50	38	-.42	0.26	0.79	-1.04	.74	-1.10	A412
Mean	54.5	47.6	.00	.24	.98	-.09	1.02	-.03
P. SD	15.6	5.2	.73	.01	.19	.93	0.36	1.11

A Principal Component Analysis of Residuals (PCAR) confirmed unidimensionality for all categories, with eigenvalues of the first contrast below 2.0. Only the total scale showed a slightly elevated value (2.73), prompting further inspection. While three clusters of items were identified based on PCAR loadings, high inter-cluster correlations ($r = 1.000$) indicated measurement of the same latent trait. Thus, the hypothesis of multidimensionality was rejected.

Additionally, local independence was confirmed: no residual correlations exceeded the critical interval of $[-0.7; 0.7]$ The correlation table (Table 4) of category A2 is shown as an example. The items A205 and A208 also had the maximum correlation across all categories with $|-.51| = .51$.

Tab. 4. Category A2 - Largest Standardized Residual Correlations

Correlation	Item 1	Item 2
-0.51	A205	A208
-0.37	A203	A206
-0.36	A209	A212
-0.35	A208	A212
-0.34	A207	A211

Following the examination of stochastic independence, the item separation indices and the item reliability values of the individual categories as well as of the overall instrument were investigated. This was carried out for both the individual categories and the overall instrument. The corresponding values are provided in Table 5. Upon closer examination, it is evident that the reliability of categories A1, A2, A3, A4, A5, and for the groupings of the entire instrument A1-A5 and A1-A6, is greater than 0.8, and thus lies at an acceptable level. Only for category A6 did the reliability fall below the necessary threshold. The quantitatively different performance between the categories A1 to A6 regarding separation can also be qualitatively seen in the Wright maps. Thus, in the Wright maps (Figs. 5 and 6), the items in category A2 cover a larger interval of personal ability than the items in category A4. This circumstance is confirmed by the bigger Item Separation of Category A2 in comparison with Category A4.

The partial credit scale was then analysed in more detail. Winstep's Table 3.2 (see table 6) was used for this. Using these tables, the steps described by Davis and Boone (2021) for analysing a multi-level scale could be carried out. Table 6 shows this for category A1.

Tab. 5. Reliability & Separation

Category	Item Separation	Item Reliability
A1	5,19	0,93
A2	3,52	0,93
A3	2,81	0,89
A4	2,31	0,84
A5	2,95	0,90
A6	1,34	0,64
A1-A5	3,01	0,90
A1-A6	2,95	0,90

Tab. 6. Partial Credit Scale - Winsteps Table 3.2 for Cateogry A1

Category Label	Observed Count	Observed Average	Outfit MNSQ	Andrich Threshold	Coherence	
					M→C	C→M
0	134	-.80	.94	NONE	64%	29%
1	238	.33	.94	-1.02	53%	73%
2	497	2.22	1.04	1.02	78%	69%

The quality of the partial credit scale was evaluated following the criteria outlined by Davis and Boone (2021), including sufficient observations per category, increasing average person ability, outfit MNSQ values within the [0.5–1.5] range, monotonic Andrich thresholds, and coherence above 40%.

For categories A1, A2, A4, and A5, all relevant parameters were within the recommended ranges. For example, in category A1, the observed response distribution, threshold spacing, and fit values confirmed the scale's sound functioning (see Table 6). In category A2, an accumulation in the "plausible" response category was noted, but coherence and fit remained acceptable. Category A4 showed slightly lower threshold distances (1.24 logits), yet the scale remained interpretable. A5 exhibited similar values to A1 and A2.

Category A3, using dichotomous items, was not included in the full partial credit analysis. However, person ability distributions and item fit values were reviewed and aligned with expectations. Category A6 displayed the weakest performance: both the coherence and fit values approached the limits of acceptability, consistent with the lower reliability values observed for this category (see Table 8).

In addition, Wright Maps were used to evaluate item targeting, comparing the distribution of item difficulties and person abilities across all categories. These maps revealed an overall good alignment between item difficulty and person ability in most categories, indicating appropriate targeting and supporting the validity of the instrument (see Table 7).

Tab. 7. Instrument's Targeting Categories.

Category	Targeting (Difference of Means)
A1-A6	0.37
A1-A5	0.41
A1	0.92
A2	0.31
A3	0.70
A4	0.53
A5	0.49
A6	0.14

Tab. 8. Summary of the Scale Analysis of Categories

Category	1.	2.	3.	4.	5.	6.	7.	8.
A1	✓	✗	✓	✓	✓	✓	✓	✓
A2	✓	✓	✓	✓	✓	✓	✓	✓
A3	✓	-	✓	✓	-	✓	-	-
A4	✓	✗	✓	✓	✓	✓	✗	✓
A5	✓	✗	✓	✓	✓	✓	✓	✓
A6	✓	✓	✓	✓	✓	✗	✓	✓

4.2 Contextual Findings

The Wright Map also allows for the analysis of individual items with reference to their content, thereby identifying patterns. Initially, we looked at the Wright Map of category A2 in figure 6 and category A4 in figure 7. Here, we could see that all items are well distributed. Item A211 appears to be the simplest, while item A212 is the most difficult one.

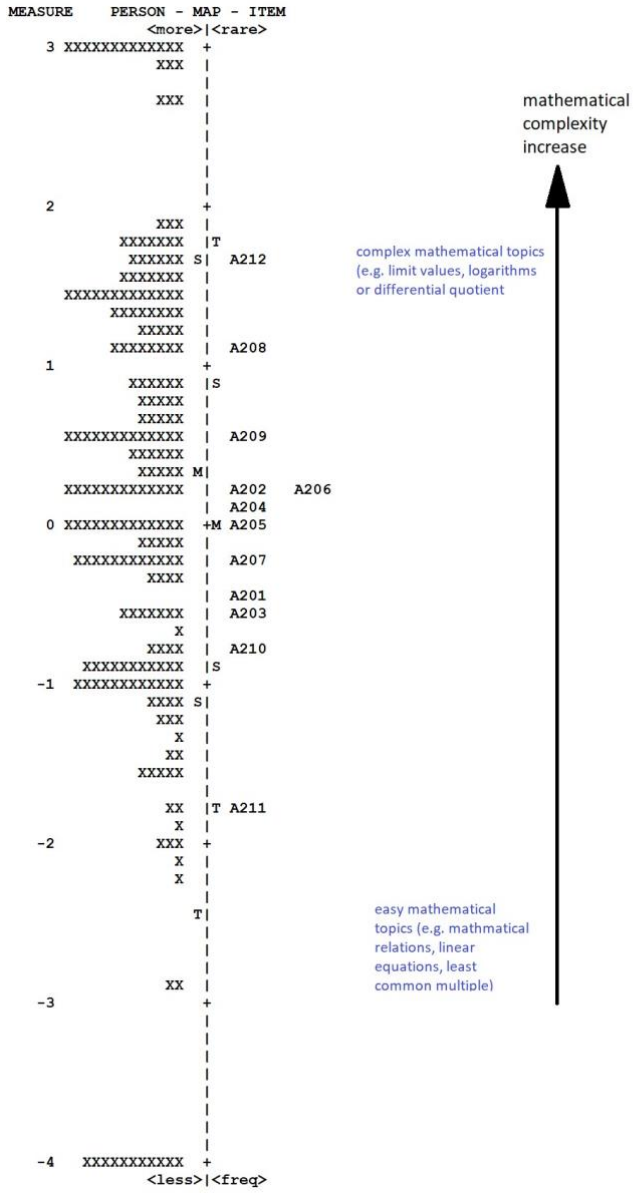

Fig. 6. Wright-Map for category A2

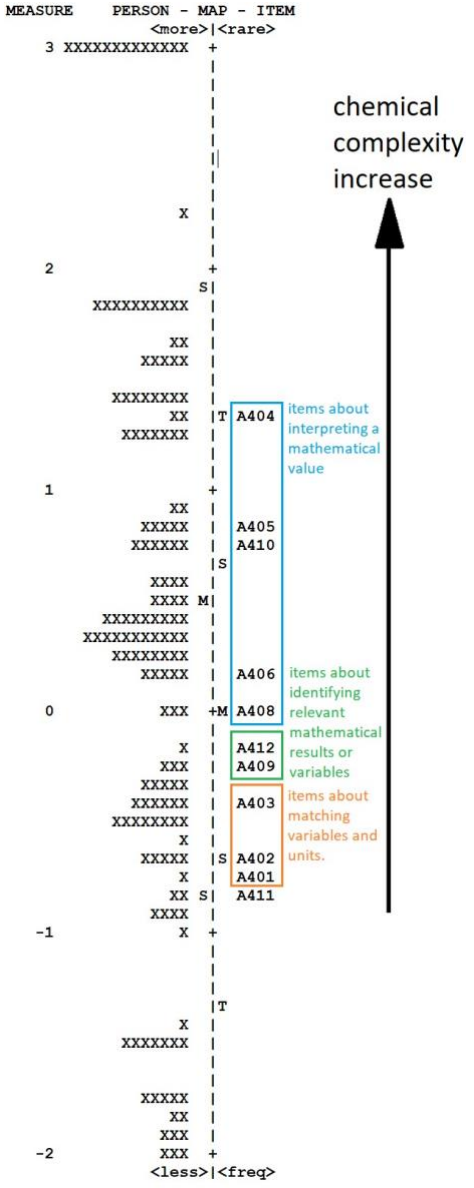

Fig. 7. Wright-Map for category A4

Since we are looking at category A2 "Mathematizing," it is worth examining the mathematics within the individual tasks and their complexity. The complexity of the mathematics is significantly influenced by the required mathematical skills and abilities. Let's consider the mathematics in four of these items. The items A208, A202, A203, and A211 differ significantly in the presentation of their answers.

Figures 8 & 9 show the translated answer formats of items A208 & A211. It is clearly visible that the answers for item A208 are given in mathematical notation, while the answers for A211 are text-based and contain no mathematical formalism. Furthermore, item A208 involves complex mathematical concepts such as the differential quotient.

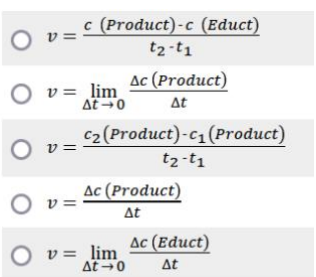

Fig. 8. Item A208 answers

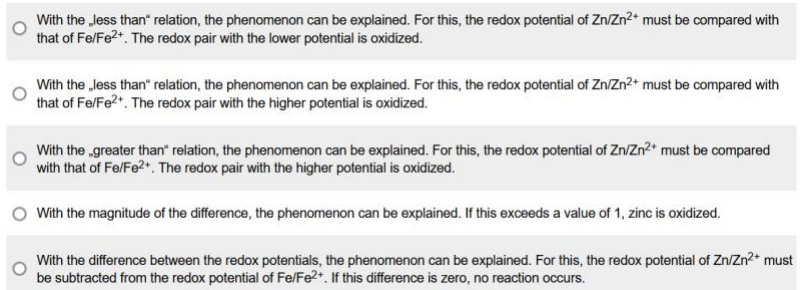

Fig. 9. Item A211 answers

In summary, for category A2, it can be noted that with increasing item difficulty, both the mathematical formalism and complexity increase. It seems that the mathematical formalism alone contributes to increasing the item difficulty. The trends are graphically represented in Figure 6. Re-examining the items of category A2, we consider the additional representations used alongside texts and symbols. In category A4, it can also be seen that, in addition to mathematical complexity, the complexity of the chemical content plays a role. The Wright Map for A4 (Fig. 7) shows that tasks requiring a chemical interpretation (e.g., A404, A405, A410) were on average more difficult than tasks where a mathematical result had to be linked with a variable or a unit had to be assigned to a mathematical result (A401, A402, A403). Additionally, it was observed that item difficulty seems to be increased by additional representations (besides text + symbol). For example, in A404, chemical structural formulas of organic compounds were presented, while in A405 only the underlying reaction equation of the considered reaction was depicted. The same was noted for items involving the identification of variables or the assignment of units. Furthermore, the difficulty was influenced not only by the task instructions but also by the content. It was shown that within the groups (interpreting, identifying & assigning) of items, difficulty increased with the growing complexity of the chemical content. The chemical complexity was thereby pegged to their respective places in the curricula. While content taught at the lower secondary level tended to reduce item difficulty, content from the upper secondary level was also difficult to interpret within the items. Here, items A401 to A403 can be exemplarily mentioned. The simplest item, A401, involved the chemical topic of "stoichiometry," item 402 the topic of "reaction kinetics," and item 403 the topic of "Michaelis-Menten kinetics." Overall, it can be held that item difficulty in category A4 increases with rising chemical complexity (operator & content).

Only one item deviates from this systematics just described. Item 411 deals with the interpretation of a mathematical relationship in the context of reaction kinetics. Similar observations can also be made for categories A1, A3, and A5. In category A1, for instance, items with image-based answer options were on average more difficult than purely text-based answer options. Category A5 revealed that especially tasks dealing with the selection of correct conclusions for a chemical result had a higher item difficulty than items where a specific given hypothesis had to be evaluated or specific conclusions for the macro-level were asked.

When all items are considered on a scale, it is evident that the individual categories have varied in difficulty for the participants. The average item difficulties for a category showed different levels of difficulty among the categories. Category A1, with an average item difficulty of −1.06 (SD = .37) logits, represents the easiest category. In particular, the mathematizing in category A2 seems to have been challenging for the participants. A2 had the highest average item difficulty with 0.22 (SD = .19). Additionally, the person abilities of the individual categories were compared, and correlations between them were investigated (See tab. 9).

Tab. 9. Correlation of Categories Person Abilities.

Category		A1	A2	A3	A4	A5
A2	Pearson Corr.	.143*				
	Sig. (2 tailed)	.024				
	N	249				
A3	Pearson Corr.	.191**	.306*			
	Sig. (2 tailed)	.006	.000			
	N	202	202			
A4	Pearson Corr.	.207**	.174*	.295**		
	Sig. (2 tailed)	.005	.020	.000		
	N	179	179	179		
A5	Pearson Corr.	.233**	.165*	.294**	.299**	
	Sig. (2 tailed)	.002	.028	.000	.000	
	N	177	177	177	177	
A6	Pearson Corr.	-.022	.038	.048	.031	-.044
	Sig. (2 tailed)	.773	.609	.517	.682	.557
	N	181	181	181	181	181
* Correlation is significant at the 0.05 level (2-tailed). ** Correlation is significant at the 0.01 level (2-tailed).						

The correlation analysis provides a detailed overview of the statistical relationships between different categories. Correlations were found between category A1 and other categories. A significant correlation with category A5 ($r = .233^{**}$, $p = .002$, $N = 177$), category A4 ($r = .207^{**}$, $p = .005$, $N = 179$), A3 ($r = .191^{**}$, $p = .006$, $N = 202$) and A2 ($r = .143^{*}$, $p = .024$, $N=249$). The results for category A2 also show a significant moderate positive correlation with A3 ($r = .306^{*}$, $p < .001$, $N = 202$), A4 ($r = .174^{*}$, $p = .020$, $N = 179$) and A5 ($r = .165^{*}$, $p = .028$, $N = 177$), suggesting that an increase in abilities in one category is associated with an increase in another. This pattern continues with significant correlations between A3 and A4 ($r = .295^{**}$, $p < .001$, $N = 179$), and between A3 and A5 ($r = .294^{**}$, $p < .001$, $N = 177$), indicating a possible common foundation of abilities in these areas.

Category A4 also significant correlation with A5 ($r = .299^{**}$, $p < .001$, $N = 1177$), supporting the hypothesis that the abilities in these categories are partially overlapping. Category A5 further shows medium positive correlations with A2 ($r = .226^{**}$, $p = .002$, $N = 188$) and A4, which strengthens the assumption of a connection between these areas of ability.

No significant correlations between A6 and the other categories were found, suggesting that the abilities in A6 might be distinct, or the relationships to other abilities are not strong enough to achieve statistical significance.

5. Discussion & Conclusions

Looking at the research question: Can Mathematical Modelling in the natural sciences be validly measured using this approach? The results provide a differentiated picture. They show that the subdivision into atomistic categories can be measured well across the entire spectrum. This is shown by the analyses of one-dimensionality (fit-statistics; PCAR). This allows the conclusion that the theoretically implied categories can also be made measurable in this way.

Specifically, this study provides insights into how students' competencies to solve real-world chemical problems are captured through a structured framework. The atomistic approach directly targets key modeling sub-skills like understanding, mathematizing, and interpreting in chemical contexts. For example, categories A1 through A5, corresponding to distinct steps within the modeling cycle, highlight how students transition between real-world problems, chemical models, and mathematical representations. This granularity enhances measurement reliability and supports the targeted development of individual competencies. By operationalizing transitions between chemistry, mathematics, and real-world contexts, the framework effectively validates the ability to measure these competencies. In contrast, the holistic approach, represented by category A6, encompasses entire modeling tasks, including metacognitive and integrative aspects of the modeling process. While this approach provides valuable insights into students' ability to handle complex, interconnected tasks, the results indicate that such tasks are less reliably measurable within the current framework. This supports findings that atomistic approaches offer greater diagnostic precision (Zöttl et al., 2011). Nevertheless, it should be noted that the larger number of items in category A3 can lead to bias effects.

Additionally, the challenges associated with category A6 indicate that holistic task types may capture additional or distinct competencies not directly aligned with mathematical modelling as operationalised here.

With the help of reliability analyses, it was also possible to determine how well the items of the individual categories/dimensions are distributed across the dimension and thus how reliably they represent it. This revealed differences in the individual categories. While the categories A1, A2, A3, A4 and A5 each had a reliability of 0.8 or greater, category A6 only achieved a reliability of .64 far below an acceptable reliability value. This result was confirmed by looking at the item separation. While the other categories had a value greater than 2.00, the item separation for category A6 was only 1.34. To summarise, the realiability analysis also showed that categories A1, A2, A3, A4 and A5 can not only be measured, but can even be measured reliably. Accordingly, it can be seen that the categories with an atomistic task approach (categories A1, A2, A3, A4, A5) can be measured reliably. Whereas category A6, which contains holistic task types, cannot be reliably measured in terms of reliability. This is consistent with the findings of Zöttl et al. (2011), who found an advantage for the atomistic approach. Compared to the existing instruments and items (Haines et al., 2001; Kaiser & Schwarz, 2006; Zöttl et al., 2011; Brand, 2014), items with a closed response format were developed here. The analysis of the partial credit scales showed that this worked for the vast majority of categories. The quality criteria for multilevel scales developed by Davis and Boone (2021) were largely fulfilled. However, it turned out that, with the exception of category A2, each category did not perform well for one or two criteria. For category A1 and A5, it was found that the distribution of participants across the response categories increases as the category increases. According to Davis and Boone (2021), the number of people in the centre of the scale should increase and decrease at the extremes. As these guidelines should also apply to rating scales with more than four response categories, this criterion should be viewed critically with regard to the three-level partial credit scale used here. The scale used here does not represent a subjective perception but merely a scale for professional correctness. For this reason, an accumulation of answers outside the centre is not problematic. It merely implies that the items in this category were answered correctly or incorrectly more frequently and thus ultimately reflects a tendency with regard to the difficulty of this category. In addition, category A6 did not fulfil the criterion of coherence from response category to person ability and vice versa. According to Davis and Boone (2021), the value of coherence should be at least 40 % in each case. This would mean that in more than 40 % of cases the category implies a certain interval of person ability and vice versa, that a certain person ability implies a certain answer category. However, (Linacre & Wright, 2000) points out that coherence, like PCAR, is only of a supporting nature as a single argument but is not sufficient to decide on the quality of a scale. In category A6, for example, the coherence for answer option 2 (’correct answer’) in the direction ’category implies personal ability’ is only 20 %. This means that only 20 % of the participants who answered the questions correctly also had the expected personal ability. This observation thus confirms the inadequate reliability of category A6 already established. Finally, category A4 showed that the distance between scale levels in category A4 was below the threshold recommended by (Davis & Boone, 2021). It is recommended that the difference between two categories should not be less than 1.4 logits. However, this is only 1.24 logits in category A4. However, if we look at the probability curves for category A4 in Figure 10, we can still identify areas on the scale for each category where one of the three response categories is most likley. The range for category 1 (’plausible’) is not quite as clear as in categories A1, A2 and A5 (see Figure 11 but much clearer than in category A6. This shows that the positive reliability also has a positive influence on the quality of the partial credit scale.

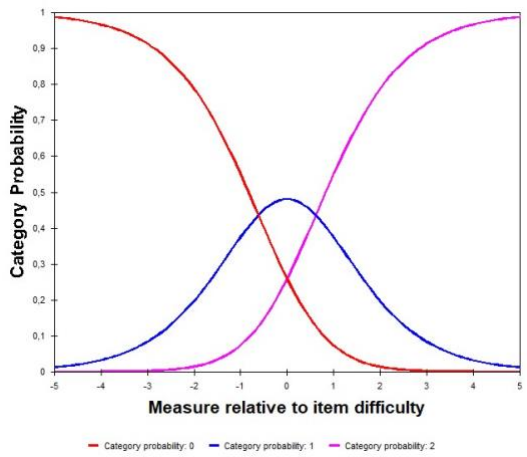

Fig. 10. Probability curves category A4

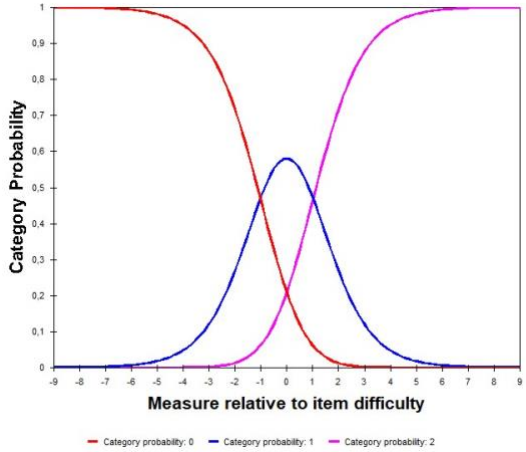

Fig. 11. Probability curves category A1

Finally, the analysis of the partial credit scale allows various conclusions to be drawn that, conceptually, a partial credit scale is a sensible approach to developing a closed response format for Mathematical Modelling. Looking at the average person ability broken down by answer category, it can be seen in all categories that this increases monotonically with increasing content correctness/quality of the answers. This in turn means that the conceptual analogy to the partial credit scales of previous studies seems to work fundamentally and that correct answers are also associated with a higher personal ability. On the other hand, however, it can also be seen that the definition and content of the individual response categories was not sufficiently clear-cut in some places due to the low level of coherence and the sometimes small differences between the response categories. Nevertheless, it can be summarised that the selected partial credit scale (with limitations) provides a suitable basis for the further development of this test option.

This limitation relates to category A6. In contrast to the other categories, this category could not be measured reliably. Accordingly, based on the quantitative results from category A6, it must be assumed that A6 captures other or additional competencies and therefore cannot contribute to the measurement of Mathematical Modelling ability. This is also confirmed by the reliability analysis: While the person abilities of categories A1 to A5 correlate significantly with each other in pairs. None of them correlate with category A6. These correlations indicate that these categories share common cognitive or conceptual foundations. The lack of significance between A6 and the other categories may indicate that A6 captures unique or less connected abilities, which could warrant special educational interventions or targeted support. These findings are important for educational practice and research as they may help to inform and refine teaching approaches and assessment strategies.

In addition, the theoretically developed categories/sub-dimensions of Mathematical Modelling also need to be critically reflected upon. If we initially limit ourselves to the Mathematical Modelling cycle according to Schmidt and Di Fuccia (2013), there are further possibilities for formulating sub-dimensions in addition to the developed sub-dimensions. The approach formulated here already delivered an unsatisfactory sub-dimension with A6 that could not be harmonised with the other subdimensions. As a result, this approach was discarded and reduced to a five-dimensional concept. Another structure that could be derived from the theory would be a two-dimensional one in which the interactions between the areas 'rest of the world', 'chemistry' and 'mathematics' are distinguished without differentiating between back and forth translation. In this case, the categories A1 and A5 as well as A2 and A4 would each fall into one category. Category A3 would represent a special case here, as it does not a priori represent any interaction between two of the areas mentioned and would therefore be omitted. However, since the ability to work mathematically has a relevance for Mathematical Modelling that should not be neglected, an approach with three sub-dimensions would also be conceivable. In this case, category A3 would be added to the two-dimensional approach already described as an additional sub-dimension.

This would address the mathematical aspect. Alternatively, steps leading to and interpreting a mathematical model could each form a sub-dimension in a three-dimensional approach.

Categories A1 and A2 would be combined into one dimension that focuses on simplifying, formalising and mathematising and categories A4 and A5 into another dimension that focuses on classifying and interpreting mathematical results. Finally, a one-dimensional structural model is also conceivable. This in turn would mean that the developed sub-dimensions do not identify themselves as such. These different structural approaches provide the basis for further analyses and studies. The proposed models can then be examined using information criteria and thus more concrete statements can be made about a possible structure of Mathematical Modelling.However, it can be said that the atomistic task approach provides a good basis for measuring Mathematical Modelling, but one that could still be improved.

The findings of this study offer valuable insights for both the design of chemistry education and the training of future chemistry teachers. The atomistic approach offers a measurable framework to identify key modeling competencies

such as understanding, mathematizing, and interpreting. These insights support teaching strategies that help students bridge theory and real-world applications.

For teacher training programs, the results highlight the potential of using the atomistic framework as a foundation for preparing educators to teach modeling skills. By emphasizing the structured progression of the modeling cycle, educators can be trained to scaffold student learning in a way that aligns with the challenges presented by both atomistic and holistic tasks. Additionally, the study underscores the need for targeted interventions to address the limitations of holistic approaches, which often involve meta-cognitive and integrative competencies. Incorporating these findings into teacher training can ensure that future educators are equipped to guide students through complex problem-solving processes and to adapt instructional methods to different levels of modeling complexity. Ultimately, these insights can contribute to refining chemistry curricula and enhancing the role of mathematical modeling in fostering scientific literacy.

References

Barke, H.-D. (2006). *Chemiedidaktik: Diagnose und Korrektur von Schülervorstellungen*. Springer.

Barke, H.-D., & Harsch, D.(2010). *Chemiedidaktik heute - Lernprozesse in Theorie und Praxis*. Springer.

Barke, H.-D., Harsch, G., Marohn, A., & Krees, S. (2015). *Chemiedidaktik kompakt*. Springer.

Blum, W., & Borromeo Ferri, R. (2009). Mathematical modelling: Can it be taught and learnt? *Journal of mathematical modelling and application, 1*(1), 45–58.

Blum, W., & Leiss, D. (2005). Modellieren im Unterricht mit der Tanken-Aufgabe. *Mathematik lehren*(128), 18–21.

Boone, W. J., & Staver, J. R. (2020). *Advances in Rasch analyses in the human sciences*. Springer.

Boone, W. J., Staver, J. R., & Yale, M. S. (2013). *Rasch analysis in the human sciences*. Springer.

Borromeo Ferri, R. (2004). *Mathematische Denkstile. Ergebnisse einer empirischen Studie*. Franzbecker. (Dissertation)

Borromeo Ferri, R. (2006). Theoretical and empirical differentiations of phases in the modelling process. *ZDM, 38*(2), 86–95.

Borromeo Ferri, R., Greefrath, G., & Kaiser, G. (eds.) (2013). *Mathematisches Modellieren für Schule und Hochschule*. Springer Spektrum.

Borromeo Ferri, R., Grünewald, S., & Kaiser, G. (2013). Effekte kurzzeitiger Interventionen auf die Entwicklung von Modellierungskompetenzen. In: R. Borromeo Ferri, G. Greefrath & G. Kaiser (eds.), *Mathematisches Modellieren für Schule und Hochschule: Theoretische und didaktische Hintergründe* (pp. 41–56). Springer Spektrum.

Brand, S. (2014). *Erwerb von Modellierungskompetenzen: Empirischer Vergleich eines holistischen und eines atomistischen Ansatzes zur Förderung von Modellierungskompetenzen*. Springer.

Cevikbas, M., Kaiser, G., & Schukajlow, S. (2022). A systematic literature review of the current discussion on mathematical modelling competencies: State-of-the-art developments in conceptualizing, measuring, and fostering. *Educational Studies in Mathematics, 109*(2), 205-236.

Czocher,A. J., Melhuish, K., Kandasamy, S. S., & Roan, E. (2021). Dual measures of mathematical modeling for engineering and other STEM undergraduates. *International Journal of Research in Undergraduate Mathematics Education, 7*, 328-350.

Davis, D. R., & Boone, W. (2021). Using Rasch analysis to evaluate the psychometric functioning of the other-directed, lighthearted, intellectual, and whimsical (OLIW) adult playfulness scale. *International Journal of Educational Research Open, 2*, 100054.

Eck, C., Garcke, H., & Knabner, P. (2017). *Mathematische Modellierung* (3rd ed.). Springer.

Fleige, J., Seegers, A., & Upmeier zu Belzen, A. (2012). *Modellkompetenz im Biologieunterricht: 7-10*. Auer.

Frey, A., Hartig, J., & Rupp, A. A. (2009). An NCME instructional module on booklet designs in large-scale assessments of student achievement: Theory and practice. *Educational measurement: issues and practice, 28*(3), 39–53.

Fu, J., & Xie, J. (2013). Comparison of mathematical modelling skills of secondary and tertiary students. In: G. Stillman, G. Kaiser, W. Blum, & J. Brown (eds.), *Teaching mathematical modelling: Connecting to research and practice* (pp. 165–173). Springer.

Göhner, M., & Krell, M. (2021). Was ist schwierig am Modellieren? Identifikation und Beschreibung von Hindernissen in Modellierungsprozessen von Lehramtsstudierenden naturwissenschaftlicher Fächer. *Zeitschrift für Didaktik der Naturwissenschaften, 27*(1), 155–180.

Goldhausen, I. (2015). *Mathematische Modelle im Chemieunterricht* (Unpublished doctoral dissertation). Universität Kassel, Berlin.

Greefrath, G., Kaiser, G., Blum, W., & Borromeo Ferri, R. (2013). Mathematisches Modellieren–eine Einführung in theoretische und didaktische Hintergründe. In: R. Borromeo Ferri, G. Greefrath & G. Kaiser (eds.),

Mathematisches Modellieren für Schule und Hochschule: Theoretische und didaktische Hintergründe (pp. 11–37). Springer Spektrum.

Grünewald, S. (2013). The development of modelling competencies by year 9 students: Effects of a modelling project. In *Teaching mathematical modelling: Connecting to research and practice* (pp. 185–194). Springer.

Haines, C., Crouch, R., & Davis, J. (2001). Understanding students' modelling skills. In: G. Stillman, G. Kaiser, W. Blum & J. Brown (eds.), *Modelling and mathematics education* (pp. 366–380). Elsevier.

Hankeln, C., Adamek, C., & Greefrath, G. (2019). Assessing sub-competencies of mathematical modelling—development of a new test instrument. In: G. Stillman. & J. Brown (eds.), *Lines of inquiry in mathematical modelling research in education* (pp. 143–160). Springer.

Hidayat, R., Adnan, M., & Abdullah, M. F. N. L. (2022). A systematic literature review of measurement of mathematical modeling in mathematics education context. *Eurasia Journal of Mathematics, Science and Technology Education, 18*(5), em2108.

Hodson, D. (1992). In search of a meaningful relationship: an exploration of some issues relating to integration in science and science education. *International Journal of Science Education, 14*(5), 541–562.

Höhner, K. (1996). Mathematisierungen im Chemieunterricht - ein Motivationshemmnis? *Zeitschrift für Didaktik der Naturwissenschaften, 2*(2), 51–71.

Houston, K., & Neill, N. (2003). Assessing modelling skills. In: S. Lamon, W. Parker, & K. Houston (eds.), *Mathematical Modelling* (pp. 155–164). Elsevier.

Izard, J. (2007). Assessing progress in mathematical modelling. In: C. Haines, P. Galbraith, W. Blum, W. & K. Sanowar (eds.), *Mathematical Modelling (ICTMA 12): Education, engineering and economics* (pp. 158–167). Horwood Publishing.

Johnstone, A. H. (1991). Why is science difficult to learn? Things are seldom what they seem. *Journal of computer assisted learning, 7*(2), 75–83.

Kaiser, G. (2007). Modelling and modelling competencies in school. In: C. Haines, P. Galbraith, W. Blum & K. Sanowar (eds.), *Mathematical Modelling (ICTMA 12): Education, engineering and economics* (pp. 110–119). Horwood Publishing.

Kaiser, G., & Schwarz, B. (2006). Mathematical modelling as bridge between school and university. *ZDM, 38*, 196–208.

Kaiser, G., & Schwarz, B. (2010). Authentic modelling problems in mathematics education—examples and experiences. *Journal für Mathematik-Didaktik, 31*(1), 51–76.

Kimpel, L. (2018). *Aufgaben in der Allgemeinen Chemie: zum Zusammenspiel von chemischem Verständnis und Rechenfähigkeit*. Logos.

Leisner, A. (2005). *Entwicklung von Modellkompetenz im Physikunterricht* (Unpublished doctoral dissertation). Universität Potsdam. (https://publishup.uni-potsdam.de/_opus4-ubp/frontdoor/deliver/index/docId/735/file/llh20_Leisner.pdf)

Linacre, J. M., & Wright, B. D. (2000). *Winsteps*. URL: http://www. winsteps. com/index. htm [accessed 2013-06-27].

Lingefjärd, T., & Holmquist, M. (2005). To assess students' attitudes, skills and competencies in mathematical modeling. Teaching Mathematics and Its Applications: *International Journal of the IMA, 24*(2-3), 123–133.

Maaß, K. (2004). Mathematisches modellieren im Unterricht? Ergebnisse einer empirischen Studie. *Journal für Mathematikdidaktik, 25*(2), 175–176.

Malec, J. F., Torsher, L. C., Dunn, W. F., Wiegmann, D. A., Arnold, J. J., Brown, D. A., & Phatak, V. (2007). The mayo high performance teamwork scale: reliability and validity for evaluating key crew resource management skills. *Simulation in Healthcare, 2*(1), 4–10.

Mayer, J. (2007). Erkenntnisgewinnung als wissenschaftliches Problemlösen. In D. Krüger & H. Vogt (eds.), *Theorien in der biologiedidaktischen Forschung* (pp. 177–186). Springer.

Meisert, A. (2008). Vom Modellwissen zum Modellverständnis–Elemente einer umfassenden Modellkompetenz und deren Fundierung durch lernerseitige Kriterien zur Klassifikation von Modellen. *Zeitschrift für Didaktik der Naturwissenschaften, 14*, 243–261.

Müller, J., Stender, A., Fleischer, J., Borowski, A., Dammann, E., Lang, M. & Fischer, H. E. (2018). Mathematisches Wissen von Studienanfanfängern und Studienerfolg. *Zeitschrift für Didaktik der Naturwissenschaften, 24*(1), 183–199.

Norris, S. P. & Phillips, L. M. (2003). How literacy in its fundamental sense is central to scientific literacy. *Science education, 87*(2), 224–240.

Pollak, H. (n.d.). The interaction between mathematics and other school subjects. In UNESCO (Ed.), *New trends in mathematics teaching* (pp. 232–248). UNESCO.

Reiners, C. (2017). *Chemie vermitteln: Fachdidaktische Grundlagen und Implikationen*. Springer-Spektrum.

Rost, J. (2004). *Lehrbuch Testtheorie-Testkonstruktion*. Huber.

Schmidt, I., & Di Fuccia, D.-S. (2013). Mathematische Modelle im Chemieunterricht. In: S. Bernholt (ed.), *Inquiry-based Learning - Forschendes Lernen* (pp. 338–340). IPN Kiel.

Stachowiak, H. (Hrsg.) (1980). *Modelle und Modelldenken im Unterricht. Anwendungen d. Allgemeinen Modelltheorie auf d. Unterrichtspraxis*. Verlag Julius Klinkhardt GmbH & Co. KG

Stachowiak, H. (1973). *Allgemeine Modelltheorie*. Springer.

Stäudel, L. (2012). Von der Modellnutzung zum Modellieren. *Naturwissenschaften im Unterricht. Chemie, 23*(132), 17–22.

Steinbruch, K. (1977). Denken in Modellen. In G. Schäfer (ed.), *Denken in Modellen* (pp. 10-17). Westermann.

Tennant, A., & Conaghan, P. G. (2007). The Rasch measurement model in rheumatology: what is it and why use it? When should it be applied, and what should one look for in a Rasch paper? *Arthritis Care & Research, 57*(8), 1358–1362.

Upmeier zu Belzen, A. (2013). Unterricht mit Modellen. In H. Gropengießer, U. Harms & U. Kattmann (eds.), *Fachdidaktik Biologie* (10th ed., pp. 325-334). Aulis.

Upmeier zu Belzen, A., & Krüger, D. (2010). Modellkompetenz im Biologieunterricht. *Zeitschrift für Didaktik der Naturwissenschaften, 16*(1), 41–57.

Wolfe, E. W., & Smith Jr, E. V. (2007). Instrument development tools and activities for measure validation using Rasch models: part II–validation activities. *Journal of applied measurement, 8*(2), 204–234.

Zech, F. (1998). *Grundkurs Mathematikdidaktik: theoretische und praktische Anleitungen für das Lehren und Lernen von Mathematik*. Beltz.

Zöttl, L., Ufer, S., & Reiss, K. (2011). Assessing modelling competencies using a multidimensional IRT approach. In: G. Kaiser, W. Blum, R. Borromeo Ferri, & G. Stillman (eds.), *Trends in Teaching and Learning of Mathematical Modelling: ICTMA14* (pp. 427–437). Springer