Analyzing student response processes to refine and validate a competency model and competency-based assessment task types

Chantal Soyka* and Niclas Schaper

Faculty of Arts and Humanities, Department of Industrial and Organizational Psychology, Paderborn University, Paderborn, Germany

Regarding competency-oriented teaching in higher education, lecturers face the challenge of employing aligned task material to develop the intended competencies. What is lacking in many disciplines are well-founded guidelines on what competencies to develop and what tasks to use to purposefully promote and assess competency development. Our work aims to create an empirically validated framework for competency-oriented assessment in the area of graphical modeling in computer science. This article reports on the use of the think-aloud method to validate a competency model and a competency-oriented task classification. For this purpose, the response processes of 15 students during the processing of different task types were evaluated with qualitative content analysis. The analysis shed light on the construct of graphical modeling competency and the cognitive demand of the task types. Evidence was found for the content and substantive aspect of construct validity but also for the need to refine the competency model and task classification.

KEYWORDS

competency-oriented assessment, task types, validation, think-aloud, competency model, graphical modeling, conceptual modeling, computer science

1 Introduction

Although in Europe the claim of educational policy for competency orientation was already initiated by the Bologna reform for more than 20 years, this approach does not seem to have been fully integrated into university teaching yet (Bachmann, 2018). It requires a new understanding of teaching in higher education and adjustments in teaching and assessment practice, which poses a challenge to many university lecturers (Schindler, 2015). This is primarily because it is not only necessary to clearly define the intended competency but also to select or develop assessment tasks that adequately foster or measure this competency. Competency orientation in the context of assessment means that assessment tasks address the intended competencies, i.e., that students are required to use relevant competency aspects (knowledge, skills, attitudes, etc.) to solve representative tasks of a domain so that valid inferences about underlying competency can be drawn from performance on the tasks (Blömeke et al., 2015; Schindler, 2015). Following Messick's (1995) unified concept of construct validity, empirical evidence of the appropriateness of the theoretical assumptions associated with the competency-based assessment and the interpretations of results must be found when

designing and validating assessments. Validity arguments for competence-oriented assessments used in the teaching-and-learning process must primarily support the claim that the tasks are suitable for supporting or measuring the development of competencies of a specific domain. Hence, based on the interpretation of students' responses inferences about their knowledge and skills should be possible, which in turn should inform instruction (Kane and Mislevy, 2017). In particular, arguments for the substantive aspect of validity are important because they provide evidence that the assessment tasks actually elicit construct-relevant cognitions or operations and thus address relevant competencies. With regard to cognitive validation, the analysis of response processes and their alignment with what is intended to be assessed is primarily advocated (e.g., Brückner and Pellegrino, 2017).

In this article, we present our approach to validating a competency model as well as a corresponding competency-oriented task classification in the field of graphical modeling in computer science based on students' response processes. The competence model specifies the competence facets relevant to the domain and serves as a theoretical basis for selecting and designing assessment tasks. The task classification defines typical tasks of the domain and the competence facets employed when solving these tasks. The validity study is intended to provide empirical evidence using the think-aloud method that, on the one hand, the construct of graphical modeling competency is appropriately and sufficiently represented in the competency model and, on the other hand, that typical assessment tasks actually address the postulated competence facets and make the knowledge and skills underlying competent performance observable and thus measurable. In this respect, we would like to use the present study to demonstrate to what extent the think-aloud method can contribute to the validation of competency models and competency-oriented tasks. Furthermore, with our research, we intend to contribute to closing the persisting gap regarding competence orientation in university teaching in the field of graphical modeling, where empirically validated didactic frameworks for competency-oriented assessment are still missing (Bogdanova and Snoeck, 2019).

2 Theoretical framework

This study draws upon and integrates competency modeling, cognitive modeling and validation approaches from educational research, which are described in this section.

2.1 The construct of competency, its modeling, and measurement

The construct under study is competency which is defined as the cognitive abilities and skills individuals have available or have learned to solve specific problems, as well as the associated motivational, volitional, and social dispositions and abilities (Weinert, 2002). It is therefore considered a multi-dimensional construct (Blömeke et al., 2015). Competency structure models consider the dimensionality and internal structure of competencies by describing the knowledge, skills, and attitudes (or competence facets) required for the domain under consideration in a differentiated and comprehensive manner (Schaper, 2009). We consider competence models to represent cognitive models

of domain mastery along the lines of Leighton (2004). Such models provide the theoretical assumptions and guidelines that can be used as a basis for the design of competency-based assessments in a specific domain (Leighton, 2004).

As competency is a latent construct, the underlying competency must be inferred from an observable performance (Blömeke et al., 2015). In university teaching, competency-oriented knowledge tests are often used, that examine latent abilities regarded as prerequisites for competent behavior (Schindler, 2015). These tests attempt to represent competency through various typical tasks, thus eliciting observable behavior from which competency can be inferred. However, it is often not possible to observe or infer all the competence facets involved in task processing based on the task solution alone (Gorin, 2006). Thus, depending on the competence facets to be examined, suitable methods or instruments must be selected that allow valid assessment.

In educational research, cognitive models of task performance are generated to define the interconnected knowledge and skills needed for solving specific tasks or responding to test items within a content domain (Leighton, 2004). Such models break down and depict the task-specific requirements and response processes. For instance, cognitive models for tasks with different difficulty levels can be generated by including appropriate hypotheses in the model, such as items with high difficulty are solved correctly by fewer students and require a longer response time than easier items (Gorin, 2006). Corresponding underlying hypotheses should also be incorporated when investigating the alignment of different task formats, such as selected-response (SR) (e.g., multiple choice questions) versus constructed-response (CR) (e.g., responses to essay questions) items (Gorin, 2006; Mo et al., 2021). If it is assumed that both response formats elicit the same response processes or address the same competencies, the cognitive models should be identical. Otherwise, specific models should be developed for the response formats. Moreover, the assessment format can also be considered in the development of cognitive models of task performance. It can be assumed that, depending on the assessment format (e.g., written or oral assessment) and whether the assessment is evaluated product-related (i.e., based on the written task solution) or process-related (i.e., based on the observation of the response process), other aspects of the examined competency can be captured (Schreiber et al., 2009). One way for developing such cognitive models of task performance is rational task analyses in which domain experts determine the task-related processes, the cognitive demand level of the tasks, or the addressed competencies (e.g., Schott and Azizi Ghanbari, 2008). Since both cognitive models of domain mastery, as well as cognitive models of task performance, are the foundation for the development and selection of proper assessment tasks and later interpretation of the results, it is essential to investigate and validate them empirically (Leighton, 2004).

2.2 Validation of competency models and interpretation of task performance

In educational research, validity is understood as "the degree to which evidence and theory support the interpretations of test scores entailed by proposed uses of tests" (American Educational Research Association et al., 2014, p. 11). Starting already in the conceptual

phase of assessment design, validation is considered an ongoing argumentative process, in which theoretical and empirical evidence is gathered to support the assumptions, inferences, and interpretations made in relation to an assessment (Kane and Mislevy, 2017). Messick (1995) distinguishes six aspects of construct validity, which serve as general validity criteria for educational assessments: content, substantive, structural, generalizability, external, and consequential. When developing competency models and competency-based assessments, evidence for the content and substantive aspects are particularly important and should be provided by appropriate empirical studies. Both aspects are linked by the core concept of representativeness (Messick, 1995). Regarding the content aspect of validity, concerning competency modeling, it must be ensured that the competency model adequately represents the specific competency domain, i.e., that it is theoretically well-founded, related to curriculum, and considered (practically) relevant (Schaper, 2017). Additionally, it should be verified that the competence facets are actually relevant and essential for the domain and that important aspects of the domain are not left out to prevent underrepresentation of the construct. Furthermore, not only should the content of the competency model and assessments be representative of the domain, but also the processes involved in task performance. This relates to the substantive aspect of validity. This aspect regards the fit between the construct and the response processes actually engaged in by test takers (Messick, 1995). According to Hubley and Zumbo (2017, p. 2) response processes are "the mechanisms that underlie what people do, think, or feel," which include behavioral and motivational aspects in addition to cognitive aspects. Empirical evidence must be found for the actual application of the intended, construct-relevant aspects of the response processes in specific test situations (Embretson and Gorin, 2001). Thus, the construct should be engaged comprehensively and realistically in performance (Messick, 1995). This can be viewed from two directions: On the one hand, the competency model, respectively the cognitive model of domain mastery, should adequately describe the construct-relevant processes that actually take place when working on domain tasks (Schaper, 2017), and on the other hand, the corresponding assessment tasks should adequately tap into the processes and facets described in the competency model, respectively, the cognitive model of task performance. Consequently, validation can be understood as confirming the alignment between the cognitive models and the actual processes involved when solving assessment tasks. A mismatch between the theoretical cognitive models and the actual cognitions stimulated by the assessment task can threaten validity (Gorin, 2006).

2.3 Graphical modeling competence and modeling task performance

Previous response process research tends to focus on psychological constructs and problem-solving processes rather than learning constructs such as subject-specific knowledge and skills (Brückner and Pellegrino, 2017). The latter is addressed in this study, as it is concerned with the construct of graphical modeling competence in computer science, its (cognitive) modeling, assessment, and validation. Graphical modeling represents a cross-cutting topic in various computing-related degree programs. It refers to the development and use of graph-based diagrams representing an existing or planned

segment of reality (e.g., in the context of a database or software design, or business process modeling) using "graphical" modeling languages, such as UML, Petri net, BPMN (Striewe et al., 2020). To define the construct of graphical modeling competency, a competency model for graphical modeling (CMGM) was developed based on learning theory, didactic assumptions, and module descriptions, and the content aspect of validity was examined and largely confirmed using an expert rating (Soyka et al., 2022). According to the CMGM, graphical modeling competency includes knowledge, skills, and affective aspects to deal with graphical models in computing-related disciplines, i.e., to understand and develop them. The competency model defines a total of 76 competence facets relevant to domain mastery along a two-dimensional matrix. On the vertical axis lies the content dimension, which distinguishes five content areas ["model understanding and interpreting" (MU), "model building and modifying" (MB), "values, attitudes, and beliefs" (VAB), "metacognitive knowledge and skills" (MC), and "social-communicative skills" (SC)]. On the horizontal axis lies the cognitive process dimension, adapted from the taxonomy according to Anderson and Krathwohl (2001), which describes four process levels ["understand" (1), "apply & transfer" (2), "analyze & evaluate" (3), and "create" (4)]. Each competence facet refers to a content area and a process level (e.g., the competence facet "MU 1.08 Learners can explain syntactic rules of the modeling language(s) under consideration." belongs to the content are "Model understanding & interpreting" and the process level "understand"). A list of all the competence facets mentioned in this article is provided as Supplementary material 1.

Furthermore, the present study draws on a classification of modeling tasks, which was developed based on a content-analytical evaluation of task material from German universities as well as a rational task analysis (Soyka et al., 2023). Modeling tasks are problem-oriented assessment tasks, each involving different competence facets. In the rational task analysis, lecturers postulated which competence facets of the CMGM are addressed by the identified task types and variants, and which content area and process level it is to be located (Soyka et al., 2023). These competency assignments to task types represent task type-specific cognitive models of task performance. As part of these cognitive models per task type, a distinction is made between process- and product-related competence facets. Product-related competence facets become evident in the behavioral product, i.e., in the final (written/documented) solution of the task. Based on the task solution, it is possible to infer and assess the underlying competence facet. Process-related competence facets become evident in the behavior during the solution process of the considered task type. It is not possible to assess these competence facets unambiguously based on the written solution or the behavioral product alone. In addition to the postulated competence facets, the cognitive models include assumptions regarding the comparability of different task variants: First, it is assumed that tasks dealing with the same content but with different response formats (SR vs. CR) generally have the same cognitive demand and are located at the same process level, thus their cognitive models are largely congruent. Second, tasks of the same task type and content but of different complexity (simple vs. complex modeling) address the same competence facets. However, time-on-task is higher and the average score achieved is lower for tasks of higher complexity. As described, the cognitive models of task performance were previously generated

by lecturers based on their assumptions and teaching experience. For this reason, empirical verification of the hypothesized competency orientation and cognitive demand is needed, particularly to gather evidence of construct validity.

3 Study objectives and research questions

The main aim of the study is to validate and confirm the theoretical cognitive models (competency model and task type classification) by investigating actual student's response processes when solving graphical modeling tasks and their alignment with the theoretical assumptions represented in the cognitive models. The underlying purpose is to gather empirical evidence for the content and substantive aspect of construct validity based on student's response processes. Furthermore, the study should provide further insights into the response processes underlying the performance on modeling tasks and thus into the construct of graphical modeling competency itself.

Specifically, with this study, we aim to answer three questions regarding valid competency modeling and assessment in graphical modeling. First, by analyzing actual response processes that occur when completing typical modeling tasks, we intend to investigate whether the competency model comprehensively and adequately represents the competence facets and response processes required for domain mastery and thus, the construct of graphical modeling competency. According to Gorin (2006, p. 23) "verbal protocols from a variety of tasks could be gathered to identify common skills that generalize to the domain as a whole." Therefore, it is hypothesized that the deeper insights into the construct reveal further relevant competence facets that have not yet been considered in the competence model. This leads to the following research question:

- Q1a: Are the actual construct-relevant response processes during the completion of typical modeling tasks comprehensively represented by corresponding competence facets and captured by the competency model for graphical modeling?

In addition, the competence facets of the competency model should adequately depict and describe the actual response processes also in terms of wording, so we pursue the following research question:

- Q1b: Are the actual response processes employed in solving typical modeling tasks adequately described and represented by the competence facets of the competency model for graphical modeling?

Second, we investigate whether the completion of typical modeling tasks involves construct-relevant response processes and adequately addresses the intended competence facets for graphical modeling. Following recommendations by Schlomske-Bodenstein et al. (n.d.), we are specifically seeking empirical evidence to validate the results and assumptions of previous rational task analysis and thus empirically investigate which competence facets are actually addressed by the task types. This aims to confirm the cognitive models of task performance by mapping them to the response processes and to provide empirical evidence for the substantive aspect of validity. This leads to the following research questions:

- Q2: Do the postulated cognitive models of task performance for typical modeling task types and variants align with the actual response processes and competency facets addressed when completing corresponding assessment tasks?

In this context, we would like to verify the assumptions regarding the comparability of different task variants, by examining the alignment between response processes across task types (cf., e.g., Mo et al., 2021), which are presented in the previous section:

- Q2a: Are tasks with the same task content (here: the same graphical model) but with different response formats (SR vs. CR) located at the same process level ("understand"), i.e., are the cognitive models of task performance largely identical as postulated?
- Q2b: Do tasks of the same task type but with different difficulty levels (here: simple vs. complex modeling) address the same competencies and can the cognitive models be confirmed?

Third, we examine whether a valid competency-based assessment is possible with the typical written task format and assessment mode. With the help of the think-aloud method, both process-related and product-related competence facets can be identified when analyzing which competence facets are addressed when working on modeling tasks. However, as written assessment tasks are considered here, the competence facets relevant to the task solution should also be reflected in the written solution so that the underlying competence facet can be inferred based on the performance in the task and appropriate feedback can be provided to learners to promote learning. This leads to the third research question:

- Q3: Can the relevant competence facets addressed by the task types be assessed based on the written task solution alone (product-related) or should the solution process also be observed for a more valid assessment (process-related)?

4 Materials and methods

The choice of study design and methods is primarily determined by the research aim. The design of the present study is of a qualitative and confirmatory nature and combines existing methods in order to adequately answer to research questions. Figure 1 provides an overview of the methods selected for data collection, analysis and interpretation. The rationale and their application are explained in more detail below.

4.1 Research method

As a method to gather empirical evidence for the substantial aspect of validity, American Educational Research Association et al. (2014) recommend the analysis of verbal response data. According to Leighton (2017), protocol analysis represents a procedure for collecting, analyzing, and interpreting this kind of data with the aim to confirm and revise existing theoretical cognitive models. Therefore, it was chosen as the research method for this study. In addition, protocol analysis is suitable for the investigation of problem-solving processes (Leighton, 2017), which are the focus of this study, since on the one hand the underlying competencies manifest themselves in problem solving and on the other hand the modeling tasks involve solving unstructured problems (Bogdanova and Snoeck, 2019).

FIGURE 1
Study design and methods.

4.2 Sampling and participants

The sampling method used can be characterized to some extent as purposive sampling (Palinkas et al., 2015), as the participants were selected both on the basis of theory and with regard to variation. First of all, since the purpose of the study was to investigate students' response processes when completing modeling tasks, the main selection criteria were that the participants were students and had prior knowledge of graphical modeling and specific modeling languages (UML and Petri nets). Thus, in accordance to theory-based sampling (Palinkas et al., 2015), the aim was to find and examine manifestations of the construct of graphical modeling competence in the verbal reports. In addition, the sample should exhibit a relatively high degree of variance in terms of learning background, knowledge and skills, in order to reflect a variation in possible response processes. Nevertheless, cost–benefit aspects also had to be considered when recruiting the participants, which means that the sample size should be sufficiently large to provide information to answer the research question but at the same time is manageable by one researcher. Additionally, students' availability and willingness to participate was a further critical factor.

Hence, we approached students from three German universities (Karlsruhe Institute of Technology, University of Duisburg-Essen, and Saarland University) via circular emails within modeling courses and contacted students directly who were known to the researchers to meet the selection criteria. This ensured that the students did not come from a single student cohort but had different backgrounds (i.e., attending different universities and modeling courses, range of knowledge, field of study). After the first wave of data collection and initial analyses, it appeared that the data collected was extensive enough to provide information to answer the research question. Fifteen students participated in the think-aloud study. Of the

participants, 12 were male and three were female. Six participants studied business informatics, three computer science, three industrial engineering, two business administration, and one software engineering. Five participants were enrolled in a bachelor's program and 10 in a master's program and were, on average, in their ninth semester ($m = 9.07$, sd $= 3.127$), and had already taken four courses related to modeling ($m = 3.93$, sd $= 1.335$, min $= 2$, max $= 6$).

Ethical approval by an independent committee was not obtained for the study because it was neither required by the project executing organization nor was it standard practice at the authors' research institution by the time the research was conducted. Nevertheless, we made every effort to comply with the ethical guidelines of the German psychological association. The participants were informed about the aim and purpose of the study and about the handling of the audiovisual and personal data collected, and consent was obtained. Participation was voluntary, independent of participation in a specific course, and incentivized with € 20. Non-participating students had no disadvantages to fear, especially not in student matters. The research involved customary classroom activities and did not deviate significantly from standard educational practices.

4.3 Selection and development of task material

Four different types of modeling tasks that address competence facets from different content areas and cognitive process levels according to the CMGM and the modeling task classification were investigated in this study. These are "Interpreting model content" (MU1/CR and MU1/SR), "Error finding" (MU3), "Model adjusting based on new requirements" (MB2), and "Model building based on a scenario" (MB4s and MB4c). In two task types, we investigated a variation of the task to verify the assumptions regarding the alignment of the corresponding cognitive models of task performance. All tasks except for the SR-task variant of the task type "Interpreting model content" are constructed response questions. A description of the task types investigated is provided as Supplementary material 2. Established tasks that had already been used in the teaching context were selected and pre-tested with two participants, which served to ensure the appropriateness and comprehensibility of the task material and task instructions as well as the usability of the interview guide.

4.4 Data collection using think-aloud procedures

In this study, the think-aloud procedure is combined with retrospective interviews. In educational research, think-aloud studies and the analysis of individual response processes (protocol analysis) are the means of choice to obtain evidence for the substantive aspect of validity (American Educational Research Association et al., 2014) and for confirming cognitive models of task performance (Leighton, 2017). In this method, students concurrently verbalize their thoughts that occur in working memory as they proceed through an assessment task (Leighton, 2017). In this way, unobservable processes and aspects of the construct become visible and analyzable, and thus evidence can be found of what a test actually measures (Leighton, 2021). This method thus allows one to investigate not only the task solution but

also the information processing when working on a task. Studies show that the concurrent verbal report method does not measurably influence the performance of the subjects (Fox et al., 2011). This was ensured in this study by not instructing the subjects to explain their thoughts during the think-aloud phase and by the instructor not intervening in the though process by asking questions. Leighton (2004) recommends a combination of the think-aloud method with retrospective interviews, in which subjects are asked about their cognitive processes, strategies used, metacognitive knowledge, or understanding only after they have completed tasks and by drawing on their long-term memory to provide a better understanding of the think-aloud phase. Furthermore, by looking at specific but typical assessment tasks, think-aloud studies can provide deeper insights into the construct under investigation and what the construct entails (Leighton, 2021). Thus, they could also serve to verify the appropriateness of the cognitive model of domain mastery, and inform theory and model building (Leighton, 2004; Gorin, 2006). Hence, additional empirical evidence can be gathered for the content aspect of construct validity.

Fifteen one-to-one sessions were conducted, in which each subject had to solve three different tasks. Task types and orders were systematically randomized. Each session was held in a seminar room, where only the subject and the instructor were present. The interview was recorded with a camcorder and a document camera to capture the editing and writing process in detail. No time requirements were imposed. It took students an average of 47.64 min to complete a set of three modeling tasks (min = 27.16 min, max = 83.25 min). To standardize the process and instructions during the sessions, an instruction guide was developed based on the recommendations of Leighton (2017). The aim, method, and procedure of the study were explained and any questions were clarified. The subjects were informed that participation was voluntary and that they could end the session at any time. The instruction was followed by the think-aloud phase. The instructor sat quietly and at some distance from the participant, reminding him/her to think aloud during extended silent periods and responding only to organizational questions. Directly after each task, a short retrospective interview was conducted, in which subjects were asked to briefly describe their solution process. In addition, questions were asked about noticeable aspects during the think-aloud phase, the perceived level of difficulty, and familiarity with the task type. This is followed by the second and third tasks analogously. Finally, a demographic questionnaire was completed with questions on age, sex (male, female, and other), study program, and prior knowledge of modeling.

4.5 Data analysis using qualitative content analysis

The transcripts of the think-aloud process and the retrospective interview, as well as the subject's task solution, served as the analysis material for our study. Spoken words, as well as relevant actions (i.e., reading, writing, drawing, silent thinking, etc.), were transcribed. Thus, both process-related (i.e., verbal reports, editing process) as well as product-related (i.e., written task solution) data is considered. The analysis material includes a total of 45 transcripts. The task solutions were scored by the lecturers, who created the tasks using scoring rubrics for the specific task types. The transcripts were analyzed

according to qualitative structuring content analysis (Mayring, 2022) using the software MAXQDA 2022.

4.5.1 The category system

The key instrument for the content analysis is a category system with a corresponding coding guide for each task type and variant. The coding guide is provided as Supplementary material 3. The category system and coding guide were developed deductively based on the CMGM, the previous rational task analysis, and theoretical considerations and adapted inductively. The category system contains two main category groups: Construct-relevant factors (A) and construct-irrelevant factors (B). The present analysis focuses on category group A, which analyzes the competence facets underlying the response processes and task solutions. This group is specific to each task type because it contains the competence facets addressed by the task type. Category A1 includes all competence facets postulated for the task type according to the rational task analysis. Thus, it depicts the postulated cognitive model of task performance for each task type. Category A2 was added inductively to account for construct-related response processes that were not part of the postulated cognitive model of task performance. The first subcategory consists of competence facets that were already described in the CMGM but not considered relevant for the specific task according to the rational task analysis (A2.1; missing competence facets). The second subcategory consists of competence facets that were newly defined during content analysis. This category accounts for construct-relevant response processes that were not yet described in the CMGM by a competence facet (A2.2; new competence facets).

4.5.2 Data analysis process

The first analysis step aimed to capture all construct-relevant response processes adequately through corresponding competence facets (i.e., categories). This targets research question Q1. If construct-relevant response processes could not be appropriately assigned to a postulated competence facet (from category A1), it was examined whether it could be covered by an existing competence facet of the CMGM; otherwise, a new competence facet was developed. Thus, during the analysis, additional competence facets were inductively added to the category system as needed (in categories A2.1 and A2.2). This process also involved linguistically adapting individual competence facets to better align them with actual cognitive processes and operations. The inductive adaptation of the category system and further development of the coding guide was carried out by the first author with a background in higher education didactics in collaboration with two lectures in the area of modeling to reach a consensus between different perspectives. In this process, the entire analysis material was run through several times and the coding was iteratively adjusted.

To assess the reliability of the final coding, an inter-rater check was conducted. For this purpose, a student assistant, without any modeling knowledge and after being introduced to the coding guide, coded 12 transcripts (i.e., two transcripts per task type). Subsequently, the agreement with the coding of the first author was determined. Inter-rater reliability was calculated according to Brennan and Prediger (1981), indicating the percent agreement between two coders, corrected for chance. The kappa was found to be $\kappa = 0.50$ after the first run. The reason for the rather low agreement, which could be judged as "moderate" according to Landis and Koch (1977), was

investigated in detail and the coding rules were further specified. Following this, the second coder adjusted her coding with the help of the adapted coding guide. The agreement after the second run was $\kappa = 0.83$, which can be judged as "almost perfect." As a further step, the first author coded the entire material a second time several months apart to determine the agreement between two time points. The intra-rater reliability was "almost perfect" with $\kappa = 0.83$.

Ultimately, qualitatively and quantitatively data were analyzed. With regard to research question Q2, a comparison is made between the postulated cognitive model of task performance and the cognitive model modified based on the analyses. Concerning the quantitative results, it is examined how many text segments coded per category (i.e., per competence facet) and how many documents (i.e., subjects) with corresponding codes provide evidence for the involvement of the respective competence facet in each task. To answer research question Q3, the transcripts were analyzed according to whether the coded competence facet refers to written aspects of the task solution (e.g., when the respondent draws the solution or writes down his or her thoughts or answer) or whether the competence facet is evident only in the verbalizations (i.e., through utterances or actions during the solution process).

5 Results

The results related to the research questions were generated by matching the response processes with the competence facets of the CMGM and the cognitive models of task performance for the task types and variants. The main results of the analysis are depicted in Figure 2. The figure shows as columns the postulated cognitive models of task performance in comparison with the validated and modified cognitive models based on the analysis of the response processes for each task type (MU1/CR to MB4c). Hence, the figure illustrates which postulated competence facets could be confirmed (green cells) and which could not (orange cells). Also, it shows which competence facets were not postulated but could be found in the verbal reports and therefore, they were added to the validated cognitive models (blue, violet, and gray cells). In the following, the main results are explained in detail in relation to the research questions.

5.1 Results regarding the validity of the CMGM in terms of the adequacy of the cognitive model of domain mastery (Q1)

Q1a: When applying the postulated cognitive models of task performance to the think-aloud protocols during the first analysis step, it was investigated if all construct-relevant response processes could be mapped to competence facets of the CMGM. The analysis shows that the CMGM adequately represents the major part of the construct-relevant cognitive processes, as most of the response processes could be mapped to a specific competence facet of the competence model. Hence, the validated cognitive models of the different task types in Figure 2 show that the response processes could largely be assigned either to the postulated competence facets (green) or to competence facets contained in the CMGM (blue, gray) that were not postulated. Nevertheless, some success-critical response processes could not yet be represented by competence facets of the competency model, which

led to the definition of new competence facets (violet). In total, based on the analysis three new competence facets (namely, MB 3.02, MB 3.03, and MB 3.04) were identified. As seen in the validated models in Figure 2, these three new competence facets are addressed by the task types and variants MB2, MB4s, and MB4c, according to the analysis. For instance, the following construct-relevant response process while completing a "model building"-task (MB4c) could not be adequately described by an existing competence facet of the CMGM:

> *…and that aspect with the days. Yes, that's still a question now, how that's done… or because it is also later no longer used somehow. I think I would do that now simply/ I would do that I think, simply as a parameter now. (…) Yes. ((writes in the box of the class "event": "+number_days: int")) As number of days, for example (…) integer. Right, so it could also be that it should somehow be made as a subclass. But since it's not discussed any more, I don't think it's that important. That's why I'm modeling this as days, as a parameter. (F_UML20210809_P02, Pos. 154–162)*

This and other verbal reports show that selecting appropriate modeling concepts (in this example the concepts of an attribute using a "parameter" vs. inheritance using a "subclass") is a success-critical skill for building graphical models and that some students fail to do so because they either do not know which concepts exist or when which concept is appropriate. Thus, we defined the new competence facet "MB 3.02 Learners can identify modeling concepts in a problem or select appropriate modeling concepts to represent specific aspects of a scenario."

Q1b: Furthermore, when mapping the response processes to the existing competence facets it became evident that linguistic clarifications are necessary for 11 competence facets considered so that they better reflect the actual response processes when solving modeling tasks. Also, the adjustments aim to improve the comprehensibility of the competence facets so that they can be better distinguished from one another. This will be demonstrated using the competence facet "MU 1.09" as an example (see Table 1). First, the verb "interpret" was added as an operator. This is because the verbal reports show that "to explain" is not explicitly required in all task types. However, it is evident from the process-related analysis that the subjects understand the semantics of types of modeling elements (e.g., a place holding a token in a given Petri net) and are thus able to interpret them, as the following text segment shows:

> *So, right. (…) Then I would first look at where the tokens are in the Petri net. We have one at "Delivery truck is at plant A", so that's where the delivery truck is located at the moment. Barrier AB is closed, and barrier BA is open. (G_PN_20210811_P07, Pos. 16–18)*

In addition, the deeper exploration of the response processes and task demands led to the sharpening, differentiation, and consistent use of terminology in the competence facets.

5.2 Results regarding the validity of the competence-oriented task classification in terms of the adequacy of the cognitive models of task performance (Q2)

To answer the second research question, the postulated cognitive models of task performance were compared with the validated

Former competence facet	Refined competence facet
MU 1.09 Learners can _explain_ the meaning (semantics) of existing modeling elements and their relationships/connections in relation to the modeling language(s) under consideration.	MU 1.09 Learners can _interpret_ or _explain_ the meaning (semantics) of existing types of modeling elements and modeling patterns in relation to the modeling language(s) under consideration.

cognitive models generated by the analysis of the response processes. The focus of the analysis lies in the comparison of the competence facets postulated with the competence facets actually addressed by the task types. Comparing the column of the postulated cognitive model with that of the validated cognitive model for each task type in Figure 2, it can be seen that central competence facets could be confirmed for most of the examined task types but considerable adjustments are also required. Across all task types, there are competence facets that have been postulated but for which no evidence could be found within the think-aloud protocols and thus, could not be confirmed ("non-confirmed"). In addition, as discussed before, evidence was found that the respective tasks address further competence facets that were not postulated _a priori_. Thus, further competence facets are added to the validated cognitive models. These are either taken from the CMGM ("missing competence facets") or newly developed and added to the CMGM and the cognitive models of task performance ("new competence facets"). It was also found that further competence facets of the CMGM were evident in individual verbal reports ("further missing but optional competence facets"). These are process-related and mainly transversal competence facets relevant to graphical modeling such as values, attitudes, and beliefs (e.g., "VAB 1.01 Learners can explain the objectives and relevance of graphical modeling for the respective area of modeling in computer science") or metacognitive skills (MC; e.g. "MC 3.03 Learners can reflect on and evaluate their level of knowledge and skills related to graphical modeling"). They are classified as "optional," as the general relevance of these for the successful solution of the tasks cannot be adequately clarified based on the present study, partly because the evidence was found only in individual verbal reports. However, the fact that there is some evidence of their involvement in solving modeling tasks suggests that these are construct-related competence aspects that are rightly part of the CMGM.

Looking at specific task types, it was found that for the task type "Model adjusting" the postulated competence facets and cognitive processes differ significantly from those actually engaged in by the students. The task type was assumed to be on the level "apply and transfer" because parts of a graphical model are already given in this task type, and missing aspects have to be completed. However, according to the analysis this task type addresses the same competence facets as the task type "Model building" (e.g., MB4.04, MB 4.05, MB4.06) and requires the development of new model parts as well as the integration of these into the existing model. Thus, this task is more cognitively demanding than assumed, so it is now located at the process level "create."

Q2a: Furthermore, the analysis shows that the assumed task variants for interpreting model content with two different response formats (CR vs. SR) do not tap into the same competence facets. The results show that describing the model content in natural language (MU1/CR) also requires social-communicative skills (SC 2.01, SC 2.02) since students have to be able to perform the translation from technical language to a generally understandable description as is evident in this text segment:

> Well, I tried not to talk about tokens or places or transitions because of natural language and tried to describe the process a bit leaner by not writing down every state here, but only the basic states that are important. (G_PN_20210810_P06, Pos. 104–107)

Similar cognitive processes do not manifest themselves in task variant MU1/SR. Second, the verbal reports of task type MU1/SR show that the response processes and thus the cognitive model are more similar to those of the task type "Error finding" (MU3) and less similar to the actually assumed task variant MU1/CR. For task type MU1/SR, the verbal reports show that subjects match the given item statements in terms of a micro scenario with the given model and check whether the statements are true, often using a backward-thinking strategy. This corresponds to the competence facet "MU 3.04 Learners can check the semantic correctness and completeness of a model in relation to the considered scenario." Therefore, the task type is now located at the process level "analyze and evaluate" instead of "understand."

Q2b: For the other investigated task variants of the task type "Model building" (MB4s and MB4c), the results show, by comparing the validated cognitive models in Figure 2, that the tasks with different difficulty levels address the same facets. However, in task MB4c, the average time-on-task is higher (MB4s = 13.21 min, MB4c = 24.19 min), and the average score is lower (MB4s = 9.25 points, MB4c = 6 points) than in task MB4s. This suggests that task MB4c is, as intended, a task variant with higher complexity and that the postulated cognitive models of task performance could be confirmed.

5.3 Results regarding the product-related assessment of competence facets (Q3)

Finally, based on the results, it was discussed and defined which competence facets can be assessed with the written task type based on the task solution (product-related) and which are more likely to manifest themselves during the solution process (process-related). In Figure 2, product-related competence facets are depicted in bold, and process-related competence facets are in italics. Overall, we find that more involved competence facets come to light through process-related observation and the think-aloud method than through analysis of the task solution alone. Due to this extended data basis, addressed competence facets become observable, which may not be directly reflected in the written task solution. This applies, for example, to the competence facet "MC 3.01 Learners can (a) plan their course of action, (b) select appropriate strategies, and (c) monitor their progress, understanding, and problem-solving in completing modeling tasks" in the task type "Model building." In a process-related assessment, the individual steps in the process and the strategies used become evident. In the following text segment, for instance, it becomes clear that the subject is able to structure the problem and break it down into three subproblems as well as select his strategy and plan his course of action:

Task type	MU1/CR Interpreting model content (constructed response)		MU1/SR Interpreting model content (selected response)		MU3 Error finding		MB2 Model adjusting		MB4s Model building (simple)		MB4c Model building (complex)	
	postulated	validated	postulated	validated	postulated	validated	postulated	validated	postulated	validated	postulated	validated
N		7		8		7		8		7		8
Postulated competence facets	MU 1.10; SC 2.01; SC 2.02; MU 1.01 (c=0, n=0); MU 3.04 (c=0, n=0)	MU 1.10 (c=51, n=7); SC 2.01 (c=7, n=4); SC 2.02 (c=15, n=6)	MU 1.10; MU 3.04; MU 1.01 (c=0, n=0)	MU 1.10 (c=19, n=7); MU 3.04 (c=43, n=8)	MU 3.03; MU 1.08; MU 3.04; MU 1.09; MU 1.10; VAB 1.04; VAB 1.02 (c=1, n=1); MU 3.02 (c=0, n=0)	MU 3.03 (c=15, n=6); MU 1.08 (c=15, n=6); MU 3.04 (c=67, n=7); MU 1.09 (c=82, n=7); MU 1.10 (c=13, n=6); VAB 1.04 (c=7, n=5); VAB 1.02 (c=1, n=1); MU 1.10 (c=0, n=0)	MB 4.03; MB 2.04; MB 2.05; MU 3.04; VAB 1.02 (c=4, n=3); MB 3.08 (c=0, n=0); MB 2.07 (c=0, n=0); MU 1.10 (c=0, n=0)	MB 4.03 (c=8, n=8); MB 2.04 (c=128, n=8); MB 2.05 (c=48, n=8); MU 3.04 (c=31, n=8); VAB 1.02 (c=4, n=3)	MB 4.01; MB 4.04; MB 4.05; MB 4.06; MB 4.07; MB 2.04; MB 3.01; MU 3.02 (c=1, n=1); MC 2.04 (c=0, n=0)	MB 4.01 (c=7, n=7); MB 4.04 (c=11, n=7); MB 4.05 (c=9, n=7); MB 4.06 (c=19, n=5); MB 4.07 (c=1, n=1); MB 2.04 (c=117, n=7); MB 3.01 (c=12, n=7); MU 3.02 (c=1, n=1)	MB 4.01; MB 4.04; MB 4.05; MB 4.06; MB 4.07; MB 2.04; MB 3.01; MU 3.02 (c=3, n=1); MC 2.04 (c=0, n=0)	MB 4.01 (c=8, n=8); MB 4.04 (c=9, n=8); MB 4.05 (c=13, n=8); MB 4.06 (c=51, n=7); MB 4.07 (c=4, n=2); MB 2.04 (c=196, n=8); MB 3.01 (c=41, n=8); MU 3.02 (c=3, n=1)
Non-postulated competence facets		MU 1.09 (c=42, n=7); MU 1.13 (c=5, n=3); MC 3.01 (c=34, n=7); VAB 1.04 (c=3, n=3); MU 1.11 (c=2, n=1); VAB 2.02 (c=1, n=1)		MU 1.09 (c=57, n=8); MU 1.13 (c=2, n=2); MC 3.01 (c=21, n=7)		MU 1.13 (c=6, n=4); SC 2.02 (c=8, n=4); MC 3.01 (c=37, n=7); MC 3.02 (c=4, n=2); MC 3.03 (c=5, n=3); MB 3.03 (c=13, n=5)		MB 3.02 (c=5, n=2); MB 3.03 (c=116, n=8); MB 3.04 (c=31, n=7); MB 3.01 (c=13, n=7); MC 3.01 (c=68, n=8); MB 4.04 (c=13, n=8); MB 4.05 (c=13, n=8); MB 4.06 (c=13, n=7); MB 3.05 (c=22, n=7); MU 1.09 (c=14, n=7); MU 1.13 (c=1, n=1); VAB 1.04 (c=9, n=5); MC 3.03 (c=1, n=1)		MB 3.02 (c=4, n=4); MB 3.03 (c=116, n=7); MB 3.04 (c=27, n=7); MB 2.05 (c=59, n=7); MB 3.05 (c=11, n=4); MC 3.01 (c=67, n=6); MC 3.03 (c=1, n=1); VAB 1.01 (c=1, n=1); VAB 1.03 (c=1, n=1); VAB 1.04 (c=3, n=3); VAB 2.02 (c=1, n=1); SC 2.06 (c=1, n=1)		MB 3.02 (c=25, n=6); MB 3.03 (c=182, n=8); MB 3.04 (c=58, n=7); MB 2.05 (c=79, n=8); MB 3.05 (c=35, n=6); MC 3.01 (c=131, n=8); MC 3.03 (c=8, n=5); VAB 1.01 (c=2, n=1); VAB 1.02 (c=3, n=2); VAB 1.04 (c=3, n=3); MC 2.03 (c=5, n=2); SC 2.06 (c=1, n=1); MB 2.12 (c=3, n=3)
content area	MU	MU	MU	MU	MU	MU	MB	MB	MB	MB	MB	MB
process level	understand (1)	understand (1)	understand (1)	analyze & evaluate (3)	analyze & evaluate (3)	analyze & evaluate (3)	apply & transfer (2)	create (4)	create (4)	create (4)	create (4)	create (4)
average time-on-task									< MB4c	13:21 min	> MB4s	24:19 min
average score									> MB4c	9.25 / 10.5	< MB4s	6 / 12

Legend:
- ▢ Confirmed Competence Facets
- ▢ Non-confirmed Competence Facets
- ▢ Missing Competence Facets
- ▢ New Competence Facets
- ▢ Further Missing / Optional Competence Facets

FIGURE 2

Comparison of postulated and validated cognitive models of task performance per task type. Cognitive models of task performance (columns) including the addressed competence facets (e.g., MU 1.10) and underlying assumptions per task type as "postulated" in the rational task analysis compared to the "validated" cognitive models as evidenced in the think-aloud study are shown. *N*, number of subjects solving the task type; *c*, number of text segments coded with the competence facet; *n*, number of verbal reports in which the competence facet was coded; MU, Model understanding and interpreting; MB, Model building and modifying. Bold figures represent product-related competence facets; italic figures represent process-related competence facets. A list of the competence facets is provided as Supplementary material 1.

So, we have basically two things that/ or three things that are interesting. We have first the status of the airplane ((marks the corresponding text passages)) with "landed" and "ready to fly" or "flying" or I don't know yet. And second the status of the doors ((marks the corresponding text passages)), "open", "closed". And then we can do this optional ((marks the corresponding text passages)) here, you can probably add that as an extra circuit with loading and unloading at the end. Exactly, um (…) I just don't know how to structure the whole thing now. I think I'll just start with the doors. (L_PN_20210810_P04, pos. 46–54)

In addition, the think-aloud protocols and process-related analysis provide a more accurate and clearer picture of construct-related cognitions, and thus, applied knowledge and skills as well as the rationale for decisions. For instance, the following text segment indicates that the subject selects a type of modeling element ("attribute" instead of "class" for the element "comment") (competence facet MB 3.03) and that he chooses the attribute type "string" based on logical reasoning (competence facet MB 3.04). Also, he explains the rationale for his decisions in each case (competence facet MB 3.05):

The question now is, is comment a class of its own or not? Uh, it doesn't say that, it just says that there are random things like (.), for example, when they participate, who builds up, and who dismantles. So that is relatively random. Therefore, I would model it simply as an attribute (…) as a string. You can write in it whatever you want. ((writes "+ comment: string")) (F_UML20210816_P12, Pos. 234–240)

In a product-related assessment, this explanation as well as the underlying selection processes, would not be recognizable to the lecturer based on the written solution (i.e., the created graphical model) so the causes for possible errors are difficult to identify.

6 Discussion

This paper describes the use of the think-aloud method in the context of validating a competency model in terms of a cognitive model of domain mastery and the corresponding competency-based task classification specifying cognitive models for task performance. By comparing the postulated cognitive models with actual response processes the study could provide evidence for the content aspect and

the substantive aspect of validity with regard to the competency model and the task classification. As Leighton (2017) describes, based on the analysis of response processes and the confirmation of cognitive models of task performance, new insights into the construct itself and thus the cognitive model of domain mastery could be gained. We found that the competency model and competence facets, on the whole, represent the domain and cognitive processes well. Nevertheless, the competency model could be optimized based on the study results in that three new competence facets (MB 3.02, MB 3.03, and MB 3.04) had to be added, and existing competence facets had to be made linguistically more precise so that they better describe the actual response processes.

The fit between the postulated cognitive models of task performance and the actual response processes was partially confirmed, providing evidence for the substantive aspect of validity. Nevertheless, the analysis provided new insights regarding the actual cognitive demands as well as addressed competence facets of the investigated task types and variants and led to adjustments in the assignments of competence facets to task types. Based on the protocol analysis, we identified further competence facets, which were not postulated in the prior rational task analysis but are indeed relevant for solving the modeling tasks. Also, we found that for two task types, the cognitive demand levels are higher than assumed. In this context, it was shown that two assumed task variants with different response formats (SR vs. CR) do not address the same competence facets. This contradicts the current state of research, according to which CR tasks can be adequately represented by corresponding SR tasks through careful task design (e.g., Mo et al., 2021). It can be assumed that in our case, the instructions of the tasks differ too much, and thus, the focus is put on different competence facets. Finally, the results indicate that for each task type, certain competence facets can be assessed based on a written task solution (product-related). However, other competence facets that are necessary for solving the task only become evident during the solution process and cannot be clearly assessed based on the task solution alone (process-related). Thus, a process-related evaluation of task performance reveals more competence facets addressed by the task so that a clearer picture of actual knowledge and skills emerges. Summing up, the results of the study show that the think-aloud method makes a valuable contribution to the empirical investigation as well as to improving the content and substantive aspects of validity in relation to the competency model and competency-oriented task classification for graphical modeling.

Based on the study, implications can be derived regarding conducting validation studies using the think-aloud method as well as for teaching graphical modeling in higher education. In general, the study indicates that the application of a competency model to the actual response processes that occur during the completion of representative tasks allows one to evaluate the completeness of the competency model and can confirm the relevance and applicability of the identified competence facets for the respective domain. Such an analysis can thus contribute to the content aspect of validity in relation to a competency model. Furthermore, the analysis of think-aloud protocols provides empirical evidence regarding the actual cognitive demand level of task types and the addressed competence facets, which differs from the results of prior rational task analysis. Our analysis could therefore confirm that it is difficult for lecturers to define cognitive demands purely based on their assumptions. This is in line with the findings of Masapanta-Carrión and Velázquez-Iturbide (2018). For this reason, in this context, we recommend combining a rational task analysis with a cognitive task analysis using the think-aloud method to gather empirical evidence for the substantive aspect of validity.

With our empirically investigated task classification, we can give lecturers an orientation for the purposive use of competency-oriented tasks. It shows that the investigated task types as a whole are suitable for developing competence facets for graphical modeling as they properly address construct-relevant processes. A key finding of our analysis for lecturers in the area of graphical modeling is that a product-related evaluation of written task solutions cannot accurately assess all of the competence facets actually addressed by the task. This implies that lecturers must be aware that a written task format is not suitable for testing all addressed competence facets. If the lecturer intends to assess students validly concerning process-related competencies, other task formats (such as oral examinations; and written reflection tasks) may be more suitable. We think that, especially during the acquisition of competencies with process-related evaluated assessments, students can be given valuable feedback regarding applied strategies as well as the selection of design decisions. In this way, it is possible to identify partial achievements or steps that led to the correct or incorrect solution (Schreiber et al., 2009), evaluate them, and provide appropriate feedback that promotes learning. So, it might be worth considering the think-aloud method as a didactic method in modeling education.

These results must be interpreted against the limitations of the study design. First, purposive sampling was only carried out to a limited extent, as attention was paid to a heterogeneous composition of the sample, but this was not done systematically (e.g., using certain quotas or tests to determine prior knowledge). Second, as is often the case with qualitative studies, we were only able to handle a relatively small sample of 15 subjects, as the analysis of the think-aloud protocols is very labor-intensive and time-consuming. This limits the possibilities for quantitative analysis (e.g., regarding characteristics of the items and participants) as well as the generalizability of the results. It remains unclear whether data saturation has actually been reached and whether further cases would lead to new findings or categories. Against this background, a replication of the study with a larger sample and with other tasks to be investigated appears worthwhile. Nevertheless, we think that the method provided largely consistent results across students, allowing conclusions to be drawn about generally underlying cognitions. Third, the interpretation of the qualitative data material is always shaped by the research team, making consistent evaluation difficult. To make the analysis method comprehensible and transparent and to standardize it to a certain extent, a detailed coding guide was developed, tested employing an interrater check, and provided as Supplementary material.

We can make out two main aspects for further research. First, the focus of our study was to analyze the competence facets addressed by modeling tasks. Beyond that, it would be interesting to focus on the problem-solving process and the use of specific strategies in the completion of modeling tasks. This could be done, for example, as an expert-novice comparison using the think-aloud method and/or tracking of editing steps by analyzing log files in computer-based assessments (e.g., Sedrakyan et al., 2014 performed a similar analysis of modeling behavior using process mining techniques and logging functionalities). This would provide valuable information on how or which particular task-specific or metacognitive strategies are used and should be taught to students. This could lead to more granular descriptions of corresponding competency facets and refinement of the competency model. Second, further investigation is needed regarding the influence and relevance of process-related competence facets and the competence facets considered as "optional" on the task solution as well as their valid assessment. This requires larger-scale studies that allow the investigation of influences of certain aspects of competency

on task completion. In addition, measurement instruments are missing to separately assess affective components of competency in particular (e.g., motivation, values, attitudes, and beliefs) and to investigate their relevance for the solution of modeling tasks.

Data availability statement

The raw data supporting the conclusions of this article will be made available by the authors, without undue reservation.

Ethics statement

Ethical approval was not required for the studies involving humans because it was neither required by the project executing organization nor was it standard practice at the authors' research institution by the time the research was conducted. Nevertheless, we made every effort to comply with the ethical guidelines of the German psychological association. The participants were students from various universities. Participation in the study was voluntary and independent of participation in a specific course. The students therefore had no disadvantages to fear, especially not in student matters. The participants were informed about the aim and purpose of the study and about the handling of the audiovisual and personal data collected, and consent was obtained. The audio and video recordings were transcribed and anonymized and then deleted. The data does not allow any identification of the participants. Furthermore, we would like to point out that the study involved minimal risk to participants and focused on routine educational activities. The research therefore involves customary classroom activities and does not deviate significantly from standard educational practices. The studies were conducted in accordance with the local legislation and institutional requirements. The participants provided their written informed consent to participate in this study. Written informed consent was obtained from the individual(s) for the publication of any potentially identifiable images or data included in this article.

Author contributions

CS: Conceptualization, Formal analysis, Investigation, Methodology, Visualization, Writing – original draft, Writing – review & editing. NS: Conceptualization, Funding acquisition, Supervision, Writing – review & editing.

Conflict of interest

The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Publisher's note

All claims expressed in this article are solely those of the authors and do not necessarily represent those of their affiliated organizations, or those of the publisher, the editors and the reviewers. Any product that may be evaluated in this article, or claim that may be made by its manufacturer, is not guaranteed or endorsed by the publisher.

Supplementary material

The Supplementary material for this article can be found online at: https://www.frontiersin.org/articles/10.3389/feduc.2024.1397027/full#supplementary-material

References

American Educational Research Association, American Psychological Association, and National Council on Measurement in Education (2014). *Standards for educational and psychological testing*. Washington, D.C: American Educational Research Association.

Anderson, L. W., and Krathwohl, D. R. (2001). *A taxonomy for learning, teaching, and assessing: a revision of Bloom's taxonomy of educational objectives*. New York: Longman.

Bachmann, H. (2018). "Higher education teaching redefined – the shift from teaching to learning" in *Competence oriented teaching and learning in higher education: essentials*. ed. H. Bachmann (Bern: hep), 14–40.

Blömeke, S., Gustafsson, J.-E., and Shavelson, R. J. (2015). Beyond dichotomies: competence viewed as a continuum. *Z. Psychol.* 223, 3–13. doi: 10.1027/2151-2604/a000194

Bogdanova, D., and Snoeck, M. (2019). CaMeLOT: an educational framework for conceptual data modeling. *Inf. Softw. Technol.* 110, 92–107. doi: 10.1016/j.infsof.2019.02.006

Brennan, R. L., and Prediger, D. J. (1981). Coefficient kappa: some uses, misuses, and alternatives. *Educ. Psychol. Meas.* 41, 687–699. doi: 10.1177/001316448104100307

Brückner, S., and Pellegrino, J. W. (2017). "Contributions of response processes analysis to the validation of an assessment of higher education students' competence in business and economics" in *Understanding and investigating response processes in validation research*. eds. B. D. Zumbo and A. M. Hubley (Cham: Springer), 31–52.

Embretson, S., and Gorin, J. (2001). Improving construct validity with cognitive psychology principles. *J. Educ. Meas.* 38, 343–368. doi: 10.1111/j.1745-3984.2001.tb01131.x

Fox, M. C., Ericsson, K. A., and Best, R. (2011). Do procedures for verbal reporting of thinking have to be reactive? A meta-analysis and recommendations for best reporting methods. *Psychol. Bull.* 137, 316–344. doi: 10.1037/a0021663

Gorin, J. S. (2006). Test design with cognition in mind. *Educ. Meas. Issues Pract.* 25, 21–35. doi: 10.1111/j.1745-3992.2006.00076.x

Hubley, A. M., and Zumbo, B. D. (2017). "Response processes in the context of validity: setting the stage" in *Understanding and investigating response processes in validation research*. eds. B. D. Zumbo and A. M. Hubley (Cham: Springer), 1–12.

Kane, M. T., and Mislevy, R. (2017). "Validating score interpretations based on response processes" in *Validation of score meaning for the next generation of assessments: the use of response processes*. eds. K. Ercikan and J. W. Pellegrino (New York, NY: Routledge), 11–25.

Landis, J. R., and Koch, G. G. (1977). The measurement of observer agreement for categorical data. *Biometrics* 33, 159–174. doi: 10.2307/2529310

Leighton, J. P. (2004). Avoiding misconception, misuse, and missed opportunities: the collection of verbal reports in educational achievement testing. *Educ. Meas. Issues Pract.* 23, 6–15. doi: 10.1111/j.1745-3992.2004.tb00164.x

Leighton, J. P. (2017). "Collecting and analyzing verbal response process data in the service of interpretive and validity arguments" in *Validation of score meaning for the next generation of assessments: the use of response processes*. eds. K. Ercikan and J. W. Pellegrino (New York: Routledge), 25–38.

Leighton, J. P. (2021). Rethinking think-Alouds: the often-problematic collection of response process data. *Appl. Meas. Educ.* 34, 61–74. doi: 10.1080/08957347.2020.1835911

Masapanta-Carrión, S., and Velázquez-Iturbide, J. Á. (2018). "A systematic review of the use of Bloom's taxonomy in computer science education," in Proceedings of the 49th ACM Technical Symposium on Computer Science Education, eds. T. Barnes, D. Garcia, E. K. Hawthorne and M. A. Pérez-Quiñones (New York: ACM), 441–446.

Mayring, P. (2022). *Qualitative Inhaltsanalyse: Grundlagen und Techniken*. Weinheim, Germany: Beltz.

Messick, S. (1995). Validity of psychological assessment: validation of inferences from persons' responses and performances as scientific inquiry into score meaning. *Am. Psychol.* 50, 741–749. doi: 10.1037/0003-066X.50.9.741

Mo, Y., Carney, M., Cavey, L., and Totorica, T. (2021). Using think-Alouds for response process evidence of teacher attentiveness. *Appl. Meas. Educ.* 34, 10–26. doi: 10.1080/08957347.2020.1835910

Palinkas, L. A., Horwitz, S. M., Green, C. A., Wisdom, J. P., Duan, N., and Hoagwood, K. (2015). Purposeful sampling for qualitative data collection and analysis in mixed method implementation research. *Admin. Pol. Ment. Health* 42, 533–544. doi: 10.1007/s10488-013-0528-y

Schaper, N. (2009). Aufgabenfelder und Perspektiven bei der Kompetenzmodellierung und -messung in der Lehrerbildung. *Lehrerbildung auf dem Prüfstand* 2, 166–199.

Schaper, N. (2017). Why is it necessary to validate models of pedagogical competency? *GMS J. Med. Educ.* 34, 1–8. doi: 10.3205/zma001124

Schindler, C. J. (2015). *Herausforderung Prüfen: Eine fallbasierte Untersuchung der Prüfungspraxis von Hochschullehrenden im Rahmen eines Qualitätsentwicklungsprogramms. [The challenge of examining: a case-based study of higher education teachers' assessment practices in the context of a quality development program]*. Dissertation. Munich, Germany: Technical University of Munich.

Schlomske-Bodenstein, N., Strasser, A., Schindler, C., and Schulz, F. (n.d.). *Handreichungen zum kompetenzorientierten Prüfen*. Munich, Germany: Technische Universität München.

Schott, F., and Azizi Ghanbari, S. (2008). *Kompetenzdiagnostik, Kompetenzmodelle, kompetenzorientierter Unterricht: Zur Theorie und Praxis überprüfbarer Bildungsstandards*. Münster, Germany: Waxmann.

Schreiber, N., Theyßen, H., and Schecker, H. (2009). Experimentelle Kompetenz messen?! *Physik und Didaktik in Schule und Hochschule* 3, 92–101.

Sedrakyan, G., Snoeck, M., and de Weerdt, J. (2014). Process mining analysis of conceptual modeling behavior of novices – empirical study using JMermaid modeling and experimental logging environment. *Comput. Hum. Behav.* 41, 486–503. doi: 10.1016/j.chb.2014.09.054

Soyka, C., Schaper, N., Bender, E., Striewe, M., and Ullrich, M. (2022). Toward a competence model for graphical modeling. *ACM Trans. Comput. Educ.* 23, 1–30. doi: 10.1145/3567598

Soyka, C., Striewe, M., Ullrich, M., and Schaper, N. (2023). Comparison of required competences and task material in modeling education. *Enterprise Model. Inform. Syst. Archit.* 18, 1–36. doi: 10.18417/EMISA.18.7

Striewe, M., Houy, C., Rehse, J.-R., Ullrich, M., Fettke, P., Schaper, N., et al. (2020). "Towards an automated assessment of graphical (business process) modeling competences: a research agenda" in *Informatik 2020* (Bonn: Gesellschaft für Informatik (GI)), 665–670.

Weinert, F. E. (2002). "Vergleichende Leistungsmessung in Schulen - eine umstrittene Selbstverständlichkeit" in *Leistungsmessungen in Schulen*. ed. F. E. Weinert (Weinheim: Beltz), 17–31.