Chapter 7
Assessing Modeling Competence with Questionnaires

Sabrina Mathesius and Moritz Krell

7.1 Introduction

The development of abilities related to models and modeling is one goal of science education on different educational levels in various countries all over the world (e.g. Australia: VCAA, 2016; Germany: KMK, 2005; USA: NGSS Lead States, 2013). Consequently, the development and evaluation of assessment instruments focusing on the different aspects of the framework for modeling competence (FMC; Chap. 1) are one important goal of science education research (cf. Nicolaou & Constantinou, 2014). Here, different methodological approaches have been applied, ranging from performance-assessment to closed-ended tasks. This chapter aims to provide an overview of studies that have employed instruments with either open-ended tasks or closed-ended tasks as a way to elicit individuals' abilities with respect to models and modeling. The aim of the chapter is to provide researchers in science education with a summary of instruments that have been proposed for the assessment of modeling competence and to discuss the advantages and limitations of each instrument on the basis of current standards for educational assessment (cf. AERA, APA, & NCME, 2014; Kane, 2013; Shavelson, 2013).

S. Mathesius (✉) · M. Krell
Freie Universität Berlin, Berlin, Germany
e-mail: Sabrina.Mathesius@fu-berlin.de

A. Upmeier zu Belzen et al. (eds.), *Towards a Competence-Based View on Models and Modeling in Science Education*, Models and Modeling in Science Education 12, https://doi.org/10.1007/978-3-030-30255-9_7

7.2 Questionnaires as Tools for Assessing Modeling Competence

Taking into account the FMC, which includes aspects and levels as possible parts of the progression of learning, there is a need for appropriate instruments for assessing individuals' abilities with respect to models and the modeling process in science. Using such instruments as diagnostic tools can help teachers improve students' learning opportunities and makes individual support possible (cf. Oh & Oh, 2011). The development and rigorous evaluation of assessment instruments with respect to competencies as highlighted in standard documents is critically important because of the possible consequences of testing for the participants but also because it was found that teachers tend to focus on "competencies specific to assessment and testing procedures" (Osborne, 2013, p. 267) in their lessons.

Shavelson (2013) proposed an approach for assessing competencies and evaluating the quality of test scores. This approach is in line with current standards for educational assessment (cf. AERA et al., 2014) and will therefore be used to illustrate crucial aspects of the assessment of modeling competence. Shavelson (2013) conceptualized competence assessment as a triangle with the construct, observation, and interpretation as its vertices. In relation to modeling competence, this means that a clear definition of this competence (*the construct*), a thorough understanding of the nature of the data gathered with an instrument (*observation*), and legitimate inferences based on these data (*interpretation*) are necessary.

The construct: By definition, competencies are complex and latent constructs that are not directly observable; an inference from an observable performance to an individual's competence has to be made (Shavelson, 2013). Modeling competence in science education is understood as a multidimensional construct (Nicolaou & Constantinou, 2014), comprising abilities to engage in modeling practices as well as knowledge about models and the modeling process in science ("meta-modeling knowledge"). Some definitions additionally include motivational aspects (e.g. Upmeier zu Belzen & Krüger, 2010). Furthermore, meta-modeling knowledge is usually subdivided into different aspects, each including hierarchical levels of understanding (cf. Krell, Upmeier zu Belzen, & Krüger, 2014a). Typically, the following aspects are considered: Describing the extent to which a model looks like the corresponding original, explaining reasons for multiple models, judging the purpose of a model, explaining how one can test a model, and demonstrating the reasons to change a model (cf. Nicolaou & Constantinou, 2014; Gilbert & Justi, 2016; Krell, Upmeier zu Belzen, & Krüger, 2016). Consequently, researchers have to define precisely which aspect of this complex construct is to be assessed.

Observation: Observation means an individual's performance on a set of tasks, where the "universe of possible tasks and responses for observing performance, […] logically follows from the definition of the construct" (Shavelson, 2013, p. 78). However, in relation to the assessment of modeling competence, there is still a large

universe of possible tasks, containing, for example, different test formats (e.g. performance-assessment, open-ended tasks, or closed-ended tasks) and different task contexts, both of which can influence the cognitive demands of a task and, consequently, the nature of the observed performance (cf. Krell, Upmeier zu Belzen, & Krüger, 2014b; Martinez, 1999).

Interpretation: Interpretation refers to the question of the extent to which valid inferences from observed performance to (the level of) an individual's competence can be drawn (Shavelson, 2013). The interpretation of test scores, especially in relation to complex constructs such as modeling competence, means generalizing from some scores to an individual's competence. For this generalization to be valid, the tasks have to be representative of "the entire universe of tasks" that are suitable for assessing the targeted construct (Shavelson, 2013, p. 79). This is important, for example, for the operationalization of the construct: The interpretation of test scores on the basis of tasks that have been developed for assessing meta-modeling knowledge as indicators of individuals' modeling competence may be questioned because modeling competence is not only comprised of meta-modeling but also the ability to engage in modeling practices and, depending on the definition, motivational aspects (Nicolaou & Constantinou, 2014; Upmeier zu Belzen & Krüger, 2010). Hence, the evaluation of the validity of the proposed interpretation of test scores is critical and complex, and different sources of evidence are usually needed to support the claim that the proposed inferences from test scores to an individual's competence are valid (e.g. evidence based on test content, response processes, relations to other variables, or internal structure; AERA et al., 2014). This is why "the evidence required for validation is the evidence needed to evaluate the claims being made" (Kane, 2015, p. 64). Gathering evidence based on *test content* hereby means analyzing the relation between the construct and observed performance, which is often a starting point for constructing questionnaires. Sources of evidence based on *test content* often consist of expert judgments. With respect to the assessment of modeling competence, it is necessary, for example, to ask why specific test formats and task contexts have been chosen and to what extent these decisions influence the intended interpretation of the test scores (cf. Krell et al., 2014b; Martinez, 1999). Gathering evidence on the basis of *response processes* takes into account individuals' reasoning while answering the tasks in order to evaluate the extent to which the expected skills and knowledge are de facto initiated (Leighton, 2004). The sources of this process are often interviews and think-aloud protocols. Gathering evidence based on *relations to other variables* means considering relevant external variables, for example, test scores from other assessments or categorical variables such as different subsamples (known groups). Furthermore, quality criteria such as objectivity and reliability are necessary prerequisites for the valid interpretation of test scores (AERA et al., 2014), and replication studies can contribute to consolidating validity arguments (cf. Borrmann, Reinhardt, Krell, & Krüger, 2014). The current concept of validity includes aspects of reliability and fairness in testing as part of the criteria that offer evidence of a sufficient *internal structure*.

7.2.1 Aims and Procedures for Analyzing Questionnaires Designed to Assess Modeling Competence

In the following, published instruments that are used to assess modeling competence will be analyzed and discussed on the basis of the ideas about competence assessments sketched out above. The publications under consideration were selected by using the Google scholar database to search the archives of five science education journals: *Journal of Research in Science Teaching* (2016 Impact Factor 3.179), *Science Education* (2.506), *International Journal of Science and Mathematics Education* (1.474), *Research in Science Education* (1.329), and *International Journal of Science Education* (1.240). The following word combinations were used: Questionnaire AND (model(l)ing OR meta model(l)ing knowledge OR model competence OR scientific models OR models in science OR model(l)ing processes) (cf. Campbell, Oh, Maughn, Kiriazis, & Zuwallack, 2015; Nicolaou & Constantinou, 2014). In addition, reference lists of pertinent articles were searched as well as articles from key authors in the field. Only articles that explicitly described instruments that were designed to assess (aspects of) FMC in adequate detail were considered.

7.2.2 Results of the Review, or: How Is Modeling Competence Assessed in Science Education?

In the following, the identified studies are summarized on the basis of the three aspects of *the construct* (Fig. 7.1), *observation* (task context and test format; Fig. 7.2), and *interpretation* (sources of evidence; Shavelson, 2013; Fig. 7.2). In addition, sample information is provided (Fig. 7.2).

7.2.2.1 The Construct

The assessed constructs were diverse, but some aspects of meta-modeling knowledge were considered in many studies (e.g. *nature of models*, *purpose of models*; Fig. 7.1). One reason for this partial consensus regarding the assessed construct may be that many authors (e.g. Crawford & Cullin, 2005; Treagust, Chittleborough, & Mamiala, 2002; van Driel & Verloop, 1999) explicitly referred to the study by Grosslight, Unger, Jay, and Smith (1991), which can therefore be seen as seminal for research on models and modeling in science education. Nonetheless, both the abstract de-contextualized approach (Krell et al., 2014b; Sins, Savelsbergh, van Joolingen, & van Hout Wolters, 2009) and the global levels of understanding (Crawford & Cullin, 2005; Krell et al., 2014a) proposed by Grosslight et al. (1991) have been critically discussed, leading to more differentiated theoretical frameworks (e.g. Crawford & Cullin, 2005; Krell et al., 2014a).

	Study	Construct					
		NAT	ALT	PUR	TES	CHA	Others
1	Crawford & Cullin (2004)	X	X	X	X	X	
2	Crawford & Cullin (2005)	X	X	X	X	X	
3	Everett et al. (2009)	X	X	X	X	X	Understanding of models in science
4	Grünkorn et al. (2014)	X	X	X	X	X	
5	Justi & Gilbert (2005)	X	X	X	X	X	
6	Justi & van Driel (2005)	X	X	X	X	X	
7	Krell (2012)	X	X	X	X	X	
8	Krell et al. (2014a)	X	X	X	X	X	
9	Krell et al. (2014b)	X	X	X	X	X	
10	Krell et al. (2015)	X	X	X	X	X	
11	Krell & Krüger (2016)	X	X	X	X	X	
12	Lin (2014)	X	X	X	X	X	Knowledge of model functions and modeling process
13	Schwarz & White (2005)	X	X	X	X	X	Meta-modeling knowledge
14	Terzer (2013)	X	X	X	X	X	
15	Treagust et al. (2004)	X	X	X	X	X	Views of models and modeling in science
16	van der Valk et al. (2007)	X	X	X	X	X	Creativity
17	Cheng & Lin (2015)	X	X	X		X	Understanding of models in science
18	Derman & Kayacan (2017)	X	X	X		X	Understanding of models in science
19	Gobert et al. (2011)	X	X	X		X	Understanding of models in science
20	Lee (2017)	X	X	X		X	Types of representations of models
21	Treagust et al. (2002)	X	X	X		X	Understanding of models in science
22	Wei et al. (2014)	X	X	X		X	Understanding of models in science
23	Lee et al. (2017)	X	X	X			Types of representations of models
24	Chittleborough et al. (2005)	X	X		X	X	Role of models in science and learning
25	Borrmann et al. (2014)	X		X	X	X	Types of representations of models
26	Gogolin et al. (2017)	X		X	X	X	
27	Sins et al. (2009)	X		X	X	X	Epistemological understanding of models and of modeling processes
28	van Driel & Verloop (1999)	X		X	X	X	Types of representations of models
29	Danusso et al. (2010)	X		X			Knowledge about scientific models and modeling
30	Gogolin (2017)	X		X			Meta-modeling knowledge
31	Krell & Krüger (2017)			X	X	X	Meta-modeling knowledge
32	Patzke et al. (2015)			X	X	X	
33	Krell et al. (2012)			X			
34	Al-Balushi (2011)						Epistemologies about the credibility of scientific models
35	Bamberger & Davis (2013)						Modeling performances

Fig. 7.1 Constructs assessed in the reviewed studies. *NAT* Nature of models; *ALT* Alternative models; *PUR* Purpose of models; *TES* Testing models; *CHA* Changing models (cf. Krell et al., 2016)

Figure 7.1 also shows that many researchers called their construct *meta-modeling knowledge* (or similar), referring to the seminal study by Schwarz and White (2005) and highlighting the procedural role of modeling as a scientific practice (e.g. Crawford & Cullin, 2005). Others emphasized the role of models as types of scientific knowledge and called their construct, for example, an *understanding of scientific models* (e.g. Treagust et al., 2002). Some researchers included both, resulting in

Study	Observation		Interpretation	Sample
	Task context Contextualized + De-contextualized	Test format (no. of tasks)	Sources of evidence	
1	De-contextualized	O (8)	Response processes, other variables	14 P
2	De-contextualized	O (8)	Response processes, other variables, replication study (1)	17 P
3	Various models + De-contextualized	O (6), concept maps, short answer, RS (27; SUMS), yes-no tasks (9)	Interrater agreement (O: = 80 %, CM: = 97.5%), internal structure, other variables, replication study (1, 21)	>200 P
4	Biological models	O (15)	Test content, interrater agreement (.81 ≤ κ ≤ .90), internal structure	1177 S
5	Chemical models, historical models + De-contextualized	RS (74; VOMM A), O (not reported; VOMM C)	Test content, response processes, other variables	19 P (RS), 74 T (RS), 63 P (O)
6	Teaching and historical models + De-contextualized	O (not reported; VOMM C)	Response processes, internal structure, other variables, replication study (5)	5 T
7	Biological models	FC (30)	Internal structure, EAP/PV reliability = .69	901 S
8	Biological models	FC (30)	Test content, internal structure, other variables	1180 S
9	Biological models + De-contextualized	RS (35)	Test content, internal structure, EAP/PV reliability = .26-.34	1349 S
10	Biological models, chemical models, physical models + De-contextualized	FC (20), O (20, short answer)	Test content, response processes, other variables	617 S (FC), 115 S (O)
11	De-contextualized	O (5)	Interrater agreement (.63 ≤ κ ≤ .87), other variables	148 T
12	De-contextualized	RS (16); web-based	Test content, response processes, internal structure, Cronbach's α = .92, other variables	187 T
13	Various models + De-contextualized	MC (18), true-false questions, categorization task	Response processes, internal structure, Cronbach's α = .19-.80, other variables	>72 S
14	Biological models	MC (40)	Test content, response processes, internal structure, EAP/PV reliability = .28-.38, other variables	1136 S
15	Molecular representations + De-contextualized	RS (44), MC + justification (6; VOMMS)	Test content, internal structure, Cronbach's α = .69-.85 (Molecular representations), Cronbach's α = .87 (VOMMS), other variables	36 S
16	Statements about common features of models	correct-incorrect + explanation (10)	Test content	77 E
17	De-contextualized	RS (27; SUMS)	Internal structure, Cronbach's α = .72-.81, other variables, replication study (21)	402 S
18	Model examples + De-contextualized	RS (26+4; SUMS)	Internal structure, Cronbach's α = .74, other variables, replication study (21)	76 P
19	De-contextualized	RS (26; SUMS)	Internal structure, Cronbach's α = .56-.86, other variables, replication study (21)	736 S
20	Biological models + De-contextualized	RS (36); web-based	Test content, internal structure, person reliability = .84, other variables	983 S
21	De-contextualized	RS (27; SUMS)	Test content, internal structure, Cronbach's α = .71-.84	228 S
22	De-contextualized	RS (27; SUMS)	Internal structure, EAP/PV reliability = .59-.68, replication study (21)	629 S
23	Biological models	O (27); web-based, true-false-questions	Response processes, interrater agreement (.80-.83), other variables	189 S
24	De-contextualized	MC + justification (5; VOMMS)	Test content, Cronbach's α = .87, other variables, replication study (15)	275 S
25	De-contextualized	RS (32)	Internal structure, Cronbach's α = .53-.68, other variables, replication study (28)	226 T
26	Biological models	FC (36)	Response processes, other variables	107 S

Fig. 7.2 Observation (task context and test format), interpretation (sources of evidence), and sample information from the reviewed studies (*Note*: Test format: *O* open-ended tasks; *MC* multiple-choice tasks; *FC* forced-choice tasks; *RS* rating scale tasks; Sample: *S* students from different school grade levels; *P* pre-service science teachers; *T* in-service science teachers; *U* university students; *E* experts; for replication studies, the replicated study is provided in parentheses. ∗ sample size mentioned in abstract $N = 1207$)

27	De-contextualized	O (10); web-based	Interrater agreement (κ = .70), other variables	26 S
28	De-contextualized	O (7), RS (32)	Test content, internal structure, Cronbach's α = .64-.75, other variables	15 T (O), 71 T (RS)
29	Real phenomena + De-contextualized	O (3+3), MC (3)	Test content, response processes, other variables, replication study (Pintó & Gutierrez, 2005)	180+115+ 93 P
30	Biological models	FC (12), O (2)	Test content, response processes, internal structure, EAP/PV reliability$_{FC}$ = .51, EAP/PV reliability$_{O}$ = .55, other variables	382 S
31	Respondents' subject of study, one chosen model known in this subject	O (6)	Interrater agreement (.64 ≤ κ ≤ .92), other variables	184 U
32	Biological models	MC (25), O (9)	Test content, internal structure, EAP/PV = .58-.75, other variables, replication study (4, 14)	514 S
33	Biological models + De-contextualized	FC (7)	Test content	1209 S*
34	Natural entities and phenomena that are located at different points along the concrete-abstract continuum	MC (19-30); specific versions for each grade level	Test content, other variables	845 S, 108 P
35	Models (smell, evaporation, friction)	O drawing (3)	Test content, interrater agreement (>.80), other variables	65 S

Fig. 7.2 (continued)

constructs such as *views of models and modeling in science* (e.g. Treagust, Chittleborough, & Mamiala, 2004). However, a closer look at the respective studies revealed that, independent of the name of the construct, most researchers included aspects related to both modeling as a practice and models as types of knowledge in their frameworks (e.g. Crawford & Cullin, 2005; Treagust et al., 2002). Therefore, if researchers want to refer to other studies, it is critically important not to rely on the given label of the construct but to precisely examine the operationalization in terms of the assessment instrument.

It is evident that the vast majority of studies included in Fig. 7.1 are related to meta-knowledge (about models, modeling, or both) but that the elements of the practice have largely been neglected (cf. Nicolaou & Constantinou, 2014). However, this neglect may be a result of the focus of this article on written assessments with questionnaires (Chap. 6).

7.2.2.2 Observation

As one aspect of observation, the abovementioned criticism of the abstract de-contextualized approach by Grosslight et al. (1991) resulted in contextualized assessments that explicitly referred to specific models or situations (e.g. Grünkorn, Upmeier zu Belzen, & Krüger, 2014). Studies have shown that the assessment context may significantly affect respondents' answers (e.g. Al-Balushi, 2011; Krell, Upmeier zu Belzen, & Krüger, 2012). These findings suggest that it is not valid to generalize observations that are based on assessments as indicators of respondents' overall meta-modeling knowledge (or similarly named constructs; see above) as long as the effect of the included contexts is not fully understood and considered (cf. Shavelson, 2013).

As another aspect of observation, the chosen task format should be considered because it can influence the cognitive demands of an assessment (Martinez, 1999). In the studies included in Fig. 7.2, open-ended task formats were chosen most often (n = 16), followed by rating scales (n = 13), multiple-choice tasks (n = 7), and forced-choice tasks (n = 6). Some researchers combined different formats, especially open-ended and rating scale tasks. The prevalence of task formats corresponds with the popularity of established instruments. For example, many researchers adopted the "Students' Understanding of Models in Science" (SUMS) questionnaire developed by Treagust et al. (2002), which uses rating scale tasks (e.g. Gobert et al., 2011).

7.2.2.3 Interpretation

The evaluation of the validity of inferences being made is a necessary prerequisite for the interpretation of assessment observations (Shavelson, 2013), and different sources of evidence have been proposed for this reason (AERA et al., 2014; Kane, 2015). In the studies shown in Fig. 7.2, evidence based on *test content* was considered most often (n = 19), for example, by conducting expert reviews of the developed instruments and judging whether the tasks adequately represent the construct (e.g. Chittleborough, Treagust, Mamiala, & Mocerino, 2005; Lin, 2014; van der Valk, van Driel, & de Vos, 2007). In addition, it should be noted that all questionnaires of the reviewed studies are based on a theoretical framework. Evidence based on *response processes* was considered in n = 12 studies, for example, by conducting concurrent (e.g. "thinking aloud"; Gogolin et al., 2017) or retrospective interviews (Justi & Gilbert, 2005; Lin, 2014). Reliability estimates (as evidence based on *internal structure*) were provided in many studies, for example, for all proposed rating scale instruments (e.g. van Driel & Verloop, 1999). Although not always explicitly treated in this way, evidence of validity based on *relations to other variables* was provided in some studies. For example, Cheng and Lin (2015) compared students' results on the SUMS questionnaire (Treagust et al., 2002) with their science learning performance and found significant positive correlations, which can be interpreted as validity evidence because it is assumed that an epistemological understanding supports the learning of science concepts (Schwarz & White, 2005).

Another important source of evidence is the implementation of replication studies (cf. Borrmann et al., 2014). Fig. 7.2 proposes that there are four instruments that have been subjected to replication studies so far: The SUMS questionnaire (Treagust et al., 2002; replicated by, e.g. Gobert et al., 2011), the questionnaire about "Models and Modeling in Science" (van Driel & Verloop, 1999; replicated by Borrmann et al., 2014), the "My Views of Models and Modeling in Science" (VOMMS) questionnaire (Treagust et al., 2004; replicated by Chittleborough et al., 2005), and the "Views on Models and Modeling C" (VOMM C) questionnaire (Justi & Gilbert, 2005; replicated by Justi & van Driel, 2005). However, only one instrument, the SUMS questionnaire, seems to be established because it has been used in several studies so far (Fig. 7.2).

7.3 Conclusion and Discussion

As stated above, validity is a fundamental requirement for the interpretation of assessment observations (Shavelson, 2013; Kane, 2013), and it "refers to the degree to which evidence and theory support the interpretations of test scores for proposed uses of tests" (AERA et al., 2014, p. 11). Kane (2013) further argued that researchers have to critically demonstrate the validity of test interpretations on the basis of a variety of evidence, especially by considering the evidence that potentially threatens the intended interpretation (cf. *falsificationism*). On the basis of the present review, it can be concluded that there are hardly any questionnaires for the assessment of modeling competence (or selected aspects) that meet these requirements (cf. Nicolaou & Constantinou, 2014). This conclusion is in line with Osborne (2013), who offered the criticism that there is a lack of evidence supporting the validity of questionnaires for assessing scientific reasoning competencies. Thus, the community needs to put more effort into the systematic evaluation of questionnaires. Two exceptional studies can be highlighted here: The SUMS questionnaire (Treagust et al., 2002) was adopted and evaluated by different researchers, resulting in validity evidence based on samples with different educational and cultural backgrounds (Cheng & Lin, 2015; Derman & Kayacan, 2017; Everett, Otto, & Luera, 2009; Gobert et al., 2011; Treagust et al., 2002; Wei et al. 2014). Furthermore, Gogolin (2017) systematically evaluated her instrument in line with the AERA et al. (2014) standards, resulting in a forced-choice questionnaire suitable for assessing 11th- to 12th-graders' meta-modeling knowledge. However, even this instrument does not take into account the influence of different task contexts on students' responses.

As discussed above, modeling competence is conceptualized as comprising abilities to engage in modeling practices, as well as knowledge about models and the modeling process in science ("meta-modeling knowledge"). Many instruments included in this review focus on single aspects of FMC, especially on the knowledge dimension of competence, and have been developed to assess, for example, students' understanding of models in science (Treagust et al., 2002) or students' meta-modeling knowledge (Gogolin, 2017). As mentioned above, the interpretation of such test scores on the basis of such tasks as indicators of individuals' modeling competence may be questioned because modeling competence not only comprises meta-modeling knowledge but also abilities to engage in modeling practices and, depending on the definition, motivational aspects (Nicolaou & Constantinou, 2014; Upmeier zu Belzen & Krüger, 2010). Therefore, the interpretation of such test scores as indicators of individuals' modeling competence would require a powerful argument for validity about, for example, meta-modeling knowledge strongly contributing to or being a prerequisite for engaging in modeling practices. This assumption has been made in the science education literature (e.g. Schwarz & White, 2005), but the empirical evidence has shown that there might not be a coherent relation between students' meta-modeling knowledge and the quality of their modeling practices (Chap. 9). Hence, depending on the goals of research, scholars have to be cautious about which instrument they choose.

One crucial aspect that is not yet understood by the research community is the influence of different task contexts on observed test performance (Al-Balushi, 2011; Krell et al., 2014b). This fundamentally calls into question the validity of existing questionnaires because the interpretation of test scores as indicators of respondents' competence levels means generalizing from "a person's performance on a small sample of tasks [...] the level of competence in the full domain" (Shavelson, 2013, p. 80). As Shavelson (2013) further emphasized, this generalization requires that the tasks on an instrument are representative of the whole universe of tasks that are suitable for assessing the targeted construct. Therefore, as long as the research community only knows that there is an effect of task contexts on test performance but is not able to explain or predict this effect, we will not be able to claim representativity, and thus, we will not be able to make valid generalizations from test scores (Krell et al., 2014b).

Another crucial aspect that directly concerns the focus of this review on written assessments is the chosen task format. In line with the argument of test score interpretation as a generalization (Shavelson, 2013), the task format is important, too. Following the established conceptualization of modeling competence as a multidimensional construct, comprising abilities to engage in modeling practices, as well as knowledge about models and the modeling process in science, the aspect of metamodeling knowledge seems to be "over-evaluated" (Nicolaou & Constantinou, 2014, p. 72), and it makes sense to ask: To what extent is it valid to argue that modeling competence can be assessed with questionnaires at all? Hence, Nicolaou and Constantinou (2014) concluded that there is a need "for a more explicit and more coherent theoretical framework for assessing knowledge, practices and processes related to the modeling competence" (p. 72).

Finally, it is important to mention that many studies included in this review were conducted before the argument-based approach for validation had been established in science education research (AERA et al., 2014; Kane, 2013; Shavelson, 2013). Most of the scholars involved in these studies did excellent work that was in line with the current standards of test development at the time. However, from a contemporary point of view, more research is clearly necessary for developing and evaluating scales and questionnaires for the assessment of the different aspects of the FMC.

References

AERA, APA, & NCME (Eds.). (2014). *Standards for educational and psychological testing*. Washington, DC: American Educational Research Association.

Al-Balushi, S. M. (2011). Students' evaluation of the credibility of scientific models that represent natural entities and phenomena. *International Journal of Science and Mathematics Education, 9*, 571–601. https://doi.org/10.1007/s10763-010-9209-4

Bamberger, Y. M., & Davis, E. A. (2013). Middle-school science students' scientific modelling performances across content areas and within a learning progression. *International Journal of Science Education, 35*, 213–238. https://doi.org/10.1080/09500693.2011.624133

Borrmann, J. R., Reinhardt, N., Krell, M., & Krüger, D. (2014). Perspektiven von Lehrkräften über Modelle in den Naturwissenschaften: Eine generalisierende Replikationsstudie [Teachers perspectives' on models in science: A generalizing replication study]. *Erkenntnisweg Biologiedidaktik, 13*, 57–72.

Campbell, T., Oh, P. S., Maughn, M., Kiriazis, N., & Zuwallack, R. (2015). A review of modeling pedagogies: pedagogical functions, discursive acts, and technology in modeling instruction. *EURASIA Journal of Mathematics, Science & Technology Education, 11*, 159-176. https://doi.org/10.12973/eurasia.2015.1314a

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

Chittleborough, G. D., Treagust, D. F., Mamiala, T. L., & Mocerino, M. (2005). Students' perceptions of the role of models in the process of science and in the process of learning. *Research in Science & Technological Education, 23*, 195–212. https://doi.org/10.1080/02635140500266484

Crawford, B. A., & Cullin, M. J. (2004). Supporting prospective teachers' conceptions of modelling in science. *International Journal of Science Education, 26*, 1379–1401. https://doi.org/10.1080/09500690410001673775

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

Danusso, L., Testa, I., & Vicentini, M. (2010). Improving prospective teachers' knowledge about scientific models and modelling: Design and evaluation of a teacher education intervention. *International Journal of Science Education, 32*, 871–905. https://doi.org/10.1080/09500690902833221

Derman, A., & Kayacan, K. (2017). Investigation of the relationship between the views of the prospective science teachers on the nature of scientific models and their achievement on the topic of atom. *European Journal of Education Studies, 3*, 541–559.

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

Gilbert, J. K., & Justi, R. (2016). *Modelling-based teaching in science education*. Cham, Switzerland: Springer.

Gobert, J. D., O'Dwyer, L., Horwitz, P., Buckley, B. C., Levy, S. T., & Wilensky, U. (2011). Examining the relationship between students' understanding of the nature of models and conceptual learning in biology, physics, and chemistry. *International Journal of Science Education, 33*, 653–684. https://doi.org/10.1080/09500691003720671

Gogolin, S. (2017). *Diagnosing students' meta-modelling knowledge: Gathering validity evidence during test development.* Doctoral dissertation. Freie Universität, Berlin, Germany. Retrieved from http://www.diss.fu-berlin.de/diss/receive/FUDISS_thesis_000000105919

Gogolin, S., Krell, M., Lange-Schubert, K., Hartinger, A., Upmeier Zu Belzen, A., & Krüger, D. (2017). Erfassung von Modellkompetenz bei Grundschüler/innen [Assessment of elementary students' modeling competence]. In H. Giest, A. Hartinger, & S. Tänzer (Eds.), *Vielperspektivität im Sachunterricht* (pp. 108–115). Bad Heilbrunn, Germany: Klinkhardt.

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

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

Justi, R., & Gilbert, J. K. (2005). Investigating teachers' ideas about models and modelling: Some issues of authenticity. In K. Boersma, M. Goedhart, O. de Jong, & H. Eijkelhof (Eds.), *Research and the quality of science education* (pp. 325–335). Dordrecht, The Netherlands: Springer.

Justi, R., & van Driel, J. (2005). The development of science teachers' knowledge on models and modelling: Promoting, characterizing, and understanding the process. *International Journal of Science Education, 27*, 549–573. https://doi.org/10.1080/0950069042000323773

Kane, M. T. (2013). Validating the interpretations and uses of test scores. *Journal of Educational Measurement, 50*, 1–73. https://doi.org/10.1111/jedm.2013.50.issue-1

Kane, M. (2015). Validation strategies. Delineating and validating proposed interpretations and uses of test scores. In M. Raymond, S. Lane, & T. Haladyna (Eds.), *Handbook of Test Development* (pp. 64-80). New York: Routledge.

KMK (Ed.). (2005). *Bildungsstandards im Fach Biologie für den Mittleren Schulabschluss [Biology education standards for the Mittlere Schulabschluss]*. München, Neuwied: Wolters Kluwer.

Krell, M. (2012). Using polytomous IRT models to evaluate theoretical levels of understanding models and modeling in biology education. *Science Education Review Letters*. *Theoretical Letters 2012*, 1–5.

Krell, M., & Krüger, D. (2016). Testing models: A key aspect to promote teaching activities related to models and modelling in biology lessons? *Journal of Biological Education, 50*, 160–173. https://doi.org/10.1080/00219266.2015.1028570

Krell, M., & Krüger, D. (2017). University students' meta-modelling knowledge. *Research in Science & Technological Education, 35*, 261–273. https://doi.org/10.1080/02635143.2016.1274724

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

Krell, M., Upmeier zu Belzen, A., & Krüger, D. (2012). Students' understanding of the purpose of models in different biological contexts. *International Journal of Biology Education, 2*, 1–34. Retrieved from http://www.ijobed.com/2_2/Moritz-2012.pdf

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

Krell, M., Upmeier zu Belzen, A., & Krüger, D. (2014b). Context-specificities in students' understanding of models and modelling: An issue of critical importance for both assessment and teaching. In C. Constantinou, N. Papadouris, & A. Hadjigeorgiou (Eds.), *E-book proceedings of the ESERA 2013 conference: Science education research for evidence-based teaching and coherence in learning* (pp. 1024–1035). Nicosia, Cyprus: European Science Education Research Association. Retrieved from https://www.esera.org/publications/esera-conference-proceedings/esera-2013

Krell, M., Upmeier zu Belzen, A., & Krüger, D. (2016). Modellkompetenz im Biologieunterricht [Modeling competence in biology education]. In A. Sandmann, & P. Schmiemann (Eds.), *Biologiedidaktische Forschung: Schwerpunkte und Forschungsstände* (pp. 83–102). Berlin, Germany: Logos.

Lee, S. W.-Y. (2017). Identifying the item hierarchy and charting the progression across grade levels: Surveying Taiwanese students' understanding of scientific models and modeling. *International Journal of Science and Mathematics Education*. https://doi.org/10.1007/s10763-017-9854-y

Lee, S. W.-Y., Chang, H.-Y., & Wu, H.-K. (2017). Students' views of scientific models and modeling: Do representational characteristics of models and students' educational levels matter? *Research in Science Education, 47*, 305–328. https://doi.org/10.1007/s11165-015-9502-x

Leighton, J. P. (2004). Avoiding misconception, misuse, and missed opportunities. The collection of verbal reports in educational achievement testing. *Educational Measurement: Issues and Practice, 23*, 6–15. https://doi.org/10.1111/j.1745-3992.2004.tb00164.x

Lin, J.-W. (2014). Elementary school teachers' knowledge of model functions and modeling processes: A comparison of science and non-science majors. *International Journal of Science and Mathematics Education, 12*, 1197–1220. https://doi.org/10.1007/s10763-013-9446-4

Martinez, M. E. (1999). Cognition and the question of test item format. *Educational Psychologist, 34*, 207–218.

NGSS Lead States (Eds.). (2013). *Next generation science standards: For states, by states*. Washington, DC: The National Academies Press.

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

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

Osborne, J. (2013). The 21st century challenge for science education: Assessing scientific reasoning. *Thinking Skills and Creativity, 10*, 265–279. https://doi.org/10.1016/j.tsc.2013.07.006

Patzke, C., Krüger, D., & Upmeier zu Belzen, A. (2015). Entwicklung von Modellkompetenz im Längsschnitt [Development of modeling competence in a longitudinal study]. In M. Hammann, J. Mayer, & N. Wellnitz (Eds.), *Lehr- und Lernforschung in der Biologiedidaktik: Vol. 6. Theorie, Empirie & Praxis: Internationale Tagung der Fachsektion Didaktik der Biologie im VBIO, Kassel 2013* (pp. 43–58). Innsbruck, Austria: Studienverlag.

Pintó, R., & Gutierrez, R. (2005). Teachers' conceptions of scientific models. In R. Pinto & D. Couso (Eds.), *Proceedings of the 5th international ESERA conference on contributions of research to enhancing students' interest in learning science* (pp. 866–868). Barcelona, Spain: ESERA.

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

Shavelson, R. J. (2013). On an approach to testing and modeling competence. *Educational Psychologist, 48*, 73–86. https://doi.org/10.1080/00461520.2013.779483

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

Terzer, E. (2013). *Modellkompetenz im Kontext Biologieunterricht: Empirische Beschreibung von Modellkompetenz mithilfe von Multiple-Choice Items* [Modeling competence in the context of biology education]. Doctoral dissertation. Humboldt-Universität, Berlin, Germany. Retrieved from: http://edoc.hu-berlin.de/dissertationen/terzer-eva-2012-12-19/PDF/terzer.pdf

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

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

Upmeier zu Belzen, A., & Krüger, D. (2010). Modellkompetenz im Biologieunterricht [Model competence in biology education]. *Zeitschrift für Didaktik der Naturwissenschaften, 16*, 41–57.

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

Van Driel, J. H., & Verloop, N. (1999). Teachers' knowledge of models and modelling in science. *International Journal of Science Education, 21*, 1141–1153. https://doi.org/10.1080/0950069992901l0

VCAA. (2016). *Victorian certificate of education. Biology*. Melbourne, VIC: VCAA.

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