Diagnosing Students' Meta-Modelling Knowledge

Gathering Validity Evidence during Test Development

Inaugural-Dissertation
to obtain the academic degree
Doctor rerum naturalium (Dr. rer. nat.)
submitted to the Department of
Biology, Chemistry and Pharmacy
of Freie Universität Berlin

by
SARAH GOGOLIN
Berlin

2017

Diese Arbeit entstand zwischen Oktober 2012 und Oktober 2016 unter der Leitung von Prof. Dr. Dirk Krüger in der Arbeitsgruppe für Didaktik der Biologie des Instituts Biologie im Fachbereich Biologie, Chemie und Pharmazie an der Freien Universität Berlin. Die Arbeit wurde selbstständig verfasst und alle Hilfsmittel wurden entsprechend aufgeführt.

1. Gutachter	Prof. Dr. Dirk Krüger, Freie Universität Berlin
2. Gutachterin	Prof. Dr. Annette Upmeier zu Belzen, Humboldt-Universität zu Berlin

Disputation am 10.11.2017

Acknowledgment

Mein Dank gilt meiner Familie und meinen Freunden. Dirk Krüger und Annette Upmeier zu Belzen sowie den Arbeitsgruppen bin ich für die konstruktive und schöne Zeit zusammen sehr dankbar. Die folgenden Seiten sind durch und für alle Schülerinnen und Schüler, die diese Arbeit mit ihren Perspektiven gefüllt haben, entstanden.

Table of contents

Zusammenfassung

Modelle dienen einerseits als Medien, um bereits etabliertes Wissen über biologische Phänomene zu verstehen, andererseits ermöglichen sie es als Instrumente der Wissenschaft, noch unbekannte biologische Phänomene zu untersuchen (Oh & Oh, 2011). Aus didaktischer Sicht ist das Modellieren eine zentrale Arbeits- und Denkweise der Naturwissenschaften und ein wesentlicher Bestandteil einer naturwissenschaftlichen Grundbildung (KMK, 2005). Empirische Studien zeigen jedoch, dass Schüler*innen die Bedeutung von Modellen im wissenschaftlichen Erkenntnisprozess nur wenig wahrnehmen (Krell et al, 2016). Die bisher entwickelten Instrumente (u. a. Grosslight et al., 1991; Grünkorn, 2014) sind aufgrund ihrer Komplexität für einen effizienten Einsatz im Biologieunterricht durch Lehrkräfte nicht geeignet. Im vorliegenden Forschungsprojekt wurde auf der Grundlage des Kompetenzmodells der Modellkompetenz (Krell et al., 2016) ein Forced Choice Diagnoseinstrument entwickelt, welches eine valide und effiziente Diagnose von Modellverstehen in den Teilkompetenzen „Eigenschaften von Modellen“ und „Zweck von Modellen“ ermöglichen soll.

Im Rahmen der ersten globalen Forschungsfrage (RQ_{GI}) wurde die valide Interpretation der Schülerantworten in den konstruierten Forced Choice Aufgaben überprüft: Inwiefern sind die Forced Choice Aufgaben geeignet, um das Modellverstehen von Schüler*innen in den Teilkompetenzen „Eigenschaften von Modellen“ und „Zweck von Modellen“ zu diagnostizieren? Hierbei wurden Evidenzen für Validität in den Bereichen „Testinhalt“, „Antwortprozesse“, „Beziehung zu anderen Variablen“ und „Interne Struktur“ (AERA et al., 2014) gesammelt. Die meisten Evidenzen unterstützen die Interpretation der Schülerantworten. Dennoch ist zu beachten, dass die valide Interpretation der Antworten nur für die

Schüler*innen der Jahrgangsstufen zehn bis zwölf genügend durch die Evidenzdaten abgebildet wird. Des Weiteren ergibt sich aus den Untersuchungen die Forderung nach einer stärkeren Aufklärung von konstrukt-irrelevanter Varianz in den entwickelten Aufgaben zum Modellverstehen, die möglicherweise durch Informationen im Aufgabenstamm verursacht wird, sowie die Forderung nach der Analyse einer möglichen Unterrepräsentation des Konstrukts des Modellverstehens aufgrund eines Mangels an Differenzierung über Kontexte.

Durch die zweite globale Forschungsfrage (RQ_{GII}) sollte das Modellverstehen von Schüler*innen beschrieben werden: Wie häufig sind die drei Niveaus des Modellverstehens in den Teilkompetenzen „Eigenschaften von Modellen“ und „Zweck von Modellen“ in den Antworten der Schüler*innen vertreten? In der Teilkompetenz „Eigenschaften von Modellen“ wählen die Schüler*innen vor allem Perspektiven auf Niveau II, welche Modelle als idealisierte Repräsentationen eines Originals charakterisieren (Krell et al., 2016). In der Teilkompetenz „Zweck von Modellen“ wählen die Schüler*innen hauptsächlich Perspektiven auf Niveau I. Dies deutet darauf hin, dass sie den Zweck der Modelle vorrangig in der Beschreibung der entsprechenden biologischen Phänomene sehen. Zusätzlich bestärken die Ergebnisse des vorliegenden Forschungsprojekts die Annahme, dass das Modellverstehen von Schüler*innen kontextspezifisch ist, da die Häufigkeit der bevorzugten Niveaus über die einzelnen Kontexte hinweg schwankt.

Summary

Models, on one hand, aid students in their understanding, on the other, enable researchers to investigate biological phenomena (Oh & Oh, 2011). From an educational point of view, the process of thinking in and about models as a form of scientific practice can be seen as one of the essential

learning goals for science students (e. g., Germany: KMK, 2005; USA: NGSS Lead States, 2013). Existing instruments (e. g., Grosslight et al., 1991; Grünkorn, 2014) that are used to assess students' understanding of models are suitable for educational research but, due to their complexity, cannot be employed by teachers when seeking to obtain direct feedback on their students. In the present research project, we developed a forced choice task diagnostic instrument based on the 'model of model competence' (Krell et al., 2016), which was to allow for a valid and efficient diagnosis of students' meta-modelling knowledge in the aspects 'nature of models' and 'purpose of models'.

In order to provide evidence for the validity of the proposed score interpretation with the instrument, the constructed forced choice tasks were put to the test within the first global research question (RQ_{GI}): In what way do evidence and theory support the interpretation of the test scores for the intended use of diagnosing students' meta-modelling knowledge?

We gathered evidence for validity based on test content, on students' response processes, on relations to other variables and on internal structure (AERA et al., 2014). Most evidence supports the valid interpretation of the proposed scores of meta-modelling knowledge. Nevertheless, we have to advise that score interpretation is sufficiently supported only for grades ten to twelve. Furthermore, there is a demand for more research concerning construct-irrelevant variance caused by information in the task stem as well as construct underrepresentation due to a possible lack of differenciation across contexts.

The second global research question was set to provide information about students' meta-modelling knowledge (RQ_{GII}): How frequently are the three levels of understanding within the aspects 'nature of models' and 'purpose of models' represented among students?

For the aspect 'nature of models', the students mainly express perspectives on level II indicating that they understand models as idealised

representations of an original according to the ‘model of model competence’ (Krell et al., 2016). For the aspect ‘purpose of models’, on the other side, the students mainly express perspectives on level I indicating that they understand the purpose of models to be the description of phenomena. In addition to these results, we assume students’ meta-modelling knowledge to be context-specific as the frequency of the preferred levels altered across contexts.

1. Introduction

In the course of the past decades, the relevance of models in and for science was redefined (Frigg & Hartmann, 2006; Passmore, Gouvea, & Giere, 2014). Models have become a central epistemological construct in the semantic view of scientific theories (Adúriz-Bravo, 2013; Magnani, Nersessian, & Thagard, 1999) and modelling is recognised one of the main scientific inquiry methods besides comparing, experimenting, and observing (Mayer, 2007). Television programs as well as newspapers and online platforms bring models into our every-day lives. Predictions about the course of the climate change are derived from models similar to the forecast of election results. Both examples show that models can be seen as reconstructions that are inspired by theory, by data, and, not only but also, by commercial interests. It also shows that the predictions made from models do not always match empirical data.

An important goal of the educational reform in Germany is to help students develop competences to cope with real-life situations (Klieme & Hartig, 2007). Reflecting about models and modelling should be part of these competences as models belong to our every-day life (KMK, 2005; NGSS Lead States, 2013)[1]. Still, there is the question of how to arrive at building students' competences with regard to models and modelling in science. Education researchers propose the conceptualisation of theoretical models of competence followed by their operationalisation in assessment instruments and by the analysis of the diagnostic information which was produced by means of these instruments (e.g., Fleischer, Koeppen, Kenk, Klieme, & Leutner, 2013; Hartig, Klieme, & Leutner, 2008). These steps have been taken for the field of model competence. German educational

[1] KMK: Ständige Konferenz der Kultusminister der Länder in der Bundesrepublik; NGSS: Next Generation Science Standards.

standards (KMK, 2005) situate models in the section of ‘scientific inquiry’ (Mayer, 2007) and thus pinpoint models as methods of science rather than putting the focus on learning their content. Upmeier zu Belzen and Krüger (2010) proposed the ‘model of model competence’ as a conceptualisation of perspectives that scientists, teachers and students may hold towards scientific models.[2] The competence model is based on epistemological considerations about models (e.g., Giere, 2001; Mahr, 2008, 2009, 2011; Mittelstraß, 2004; Stachowiak, 1973) and on pre-existing empirically based structures of model competence (e.g., Crawford & Cullin, 2005; Grosslight, Jay, Unger, & Smith, 1991; Justi & Gilbert, 2003; Schwarz et al., 2009). It distinguishes the five aspects ‘nature of models’, ‘alternative models’, ‘purpose of models’, ‘testing models’ and ‘changing models’, while defining three levels of understanding for each aspect (Krell et al., 2016). The framework was operationalised in different assessment instruments and thus empirically evaluated (Gogolin & Krüger, 2016a; Grünkorn, 2014; Krell, 2013; Terzer, 2012; Trier, Krüger, & Upmeier zu Belzen, 2014). Finally, implications and concrete suggestions for teaching were derived from the competence model and from associated empirical findings (e.g., Fleige, Seegers, Upmeier zu Belzen, & Krüger, 2012; Gogolin & Krüger, in press; Grünkorn, Lotz, & Terzer, 2014).

The diagnostic information created by using the assessment instruments can help teachers improve and individualise their teaching about models and modelling (Hartig et al., 2008; Oh & Oh, 2011). Many researchers argue that differentiated diagnostic information is a prerequisite for the promotion of students’ model competence (e.g., Campbell, Schwarz, & Windschitl, 2016; Henze, van Driel, & Verloop, 2008; Justi & van Driel, 2005) but that teachers have trouble diagnosing their students with regard to epistemological perspectives (Aufschnaiter et al., 2012; Günther, Fleige,

[2] In this thesis, I will quote the article by Krell, Upmeier zu Belzen, and Krüger (2016) when referring to the latest published version of the ‘model of model competence’.

Upmeier zu Belzen, & Krüger, 2016; Oh & Oh, 2011). For the service of supporting teachers by providing differentiated diagnostic information to be prosperous, the instrument must be eligible for individual diagnosis (Hartig et al., 2008) and the aforementioned use of the diagnostic information must be carefully validated. Contemporary Standards for Educational and Psychological Testing (AERA, APA, & NCME, 2014)[3] set out guidelines for test construction and evaluation with the purpose of gathering evidence for validity for the proposed score interpretation. A sound validity argument is hereafter based on four pillars containing evidence based on (1) test content, (2) response processes, (3) relations to other variables and (4) internal structure (AERA et al., 2014).

In the present research project, we developed a diagnostic instrument that, on the basis of students' responses to forced choice tasks, provides immediate individual feedback about students' meta-modelling knowledge. The levels from the 'model of model competence' (Krell et al., 2016) serve as the theoretical basis for the determination of students' meta-modelling knowledge which in return is seen as an indication of students' model competence (Hartig et al., 2008; Krell, 2013). The main body of the research project was the validation of the diagnostic information which is to be derived from the students' responses to forced choice tasks. We pursued a number of studies to empirically gather evidence for validity. The majority of the findings of these studies have been published (cf. chapter 11) but ought to be summarised, integrated and discussed with reference to an overall validity argument here. In a second part of this thesis, I will delineate and discuss findings from a study with the final diagnostic instrument. Suggestions for promotional activities will be presented and a prospective for future research in the field of models and modelling in science will be pointed out.

[3] AERA: American Educational Research Association; APA: American Psychological Association; NCME: National Council on Measurement in Education.

2. Theoretical Frame and Scope

By pursuing the aim of developing a new diagnostic instrument for students' meta-modelling knowledge, one faces a number of theoretical considerations. First, one needs to grasp the content domain by defining what exactly is to be diagnosed and why. Second, one needs to evaluate and make use of the achievements of previous attempts at the same or closely related aims. Third, one needs to respect methodological instructions concerning test development and evaluation. This chapter is set to outline mayor aspects in the mentioned points.

2.1. Models and Modelling in Science

Taking a look at the research literature in the domain of models and modelling brings to light an abundance of issues worthy to be addressed. Frigg and Hartmann (2006), for example, raise questions concerning how models are connected to theory, they ask about the ontology of models and they reflect about their epistemology.

How do models relate to theory?

The philosophical debate about what models ought to be and how they are connected to theories has taken a turn in the recent years from the syntactic view towards the semantic view of theories (e.g., Adúriz-Bravo, 2013; Bailer-Jones, 1999, 2003, 2009; Magnani et al., 1999; Morgan & Morrison, 1999; Suppe, 1974; Suppes, 1961; Tarski, 1966). While in the syntactic view, a scientific model could have been any example of a theory (e.g., balls in a bottle as a model of genetic drift; cf. Frigg & Hartmann, 2006), with the

rise of the semantic view of theories in science philosophy, models are "beginning to be considered central to doing science" (Bailer-Jones, 2009, p. 127). Models rather than theories are being regarded as entities that describe empirical reality (Bailer-Jones, 2009) and thus as central units of theory formation. The semantic view derives its concept of a model from mathematical model theory where a model is a possible realisation that satisfies the theory (Bailer-Jones, 2009; Suppes, 1961; Tarski, 1966) and a theory is a family of models. Despite criticising the concept for not entirely capturing the diversity of scientific work, Suppes (1961) proposes models to be a bridge between theory and data and thus promotes them to central instruments of science. Adúriz-Bravo (2013) analyses some more recent approaches within the semantic view and sums up that Fred Suppe, Bas van Fraassen and Ron Giere "consider that theories are best identified and characterised by their corresponding classes of models; they [the researchers] therefore deem more relevant to meta-theoretically study models than theories." (p. 1604).

What kind of things are models?

When meta-theoretically studying models, one ought to ask about the very nature of models. Despite numerous attempts to answer the question "What kind of things are models?", the plurality of phenomena that may be modelled and the great spectrum of a model's use omits the formulation of a clear-cut definition (Oh & Oh, 2011). Harré (1988) concludes that „nothing is a model as such. An entity is a model only if considered in relation to something else." (p. 122). Science philosophers and science education researchers agree that models are the result of a theory-driven modelling process and are defined only in the context of their use (Bailer-Jones, 2002; Giere, 2001; Mahr, 2011; Odenbaugh, 2005; Stachowiak, 1973; van der Valk, van Driel, & Vos, 2007).

Stachowiak (1973) presents a general theory of modelling where some of the attributes from an original (or target) are being functionally mapped onto attributes of a model. His approach is primarily representational. The general theory by Stachowiak (1973) can be augmented by ideas proposed by Mahr (2008, 2009, 2011) who points out that models "do not incarnate any form of truth, but rather forms of demonstrability, possibility, and choice" (Mahr, 2011, p. 303). He bypasses the struggle for an explicit ontological definition by epistemologically determining interdependent relationships that justify something to be conceived of as a model: "Ontologically, a model is something as which something is being conceived of, and concretely, being a model is the content of a judgement in which something is being conceived of as a model." (Mahr, 2011, p. 301). Although he expects the reproach that any definition, which would be applicable to all modelled phenomena, would in fact capture nothing at all and thus be methodically useless, Mahr (2011) proposes his universal 'Epistemic Pattern of Model-Being', where he sees the question of model-being as a question of reasoning about a judgement on model-being. The pattern comprises four interdependent relationships (Figure 1).

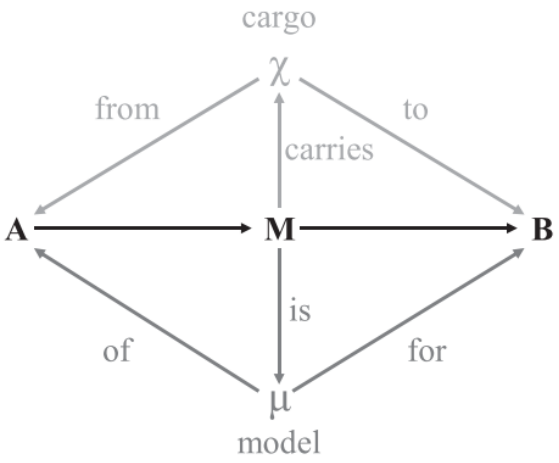


Figure 1. Epistemic Pattern of Model-Being. Conceiving of the object *M* as a model μ of *A* and for *B*, thereby carrying the cargo χ from *A* to *B* (Mahr, 2011).

The first relationship is between a model μ and the model-object *M*. This distinction is crucially important because it allows someone to perceive any object either as such or as a model-object. The second relationship focusses on the creation of the model and is to be found between the model-object *M* and that of which the model-object is conceived of as a model (model of *A*). The third relationship focusses on the model's application for a certain purpose and is in place between the model-object *M* and that for which the model-object is conceived of as a model (model of *B*). The second and the third relationship ('model of something' and 'model for something') can also be seen as epistemological functions or roles a model adopts. Finally, the fourth relationship is between the model-object *M* and the cargo χ carried by the model-object as a model. The introduction of a cargo, which may be transported better or worse, can help explain why models are being conceived differently in different contexts (Mahr & Wendler, 2009).

Just as the plurality of phenomena that may be modelled impedes a definition of the term 'model', it triggers attempts to find classification systems. While Frigg and Hartmann (2006) distinguish semantically between 'Models of Phenomena', 'Models of Data' and 'Models of Theory', Gilbert, Boulter, and Elmer (2000) specify the epistemological status of models by dividing into 'mental models', 'expressed models', 'consensus models', 'scientific models', 'historical models', 'curricular models', 'teaching models', 'hybrid models' and 'models of pedagogy'. The ontological approach of Boulter and Buckley (2000) focusses on different modes of representation (Concrete; Verbal; Visual; Mathematical; Gestural) and on attributes of representation (Quantitative vs. Qualitative; Static vs. Dynamic; Deterministic vs. Stochastic). Other ontological approaches open the dichotomy of physical to conceptual/symbolic (Coll, 2006; Suckling, Suckling, & Suckling, 1978) or pure physical to pure mathematical (Laing, 1981). The two last-named classifications compare to the results of an empirical study by Krell, Upmeier zu Belzen, and Krüger (2014b) who

developed a student-based typology of biological models and found students to distinguish between rather concrete and rather abstract models.

Which epistemological functions do models entail?

We can derive from the epistemological distinction between 'model of something' and 'model for something' that models might be used for purposes that can be assigned to representation and instrumentation. Cartier, Rudolph, and Stewart (2001) state that "scientific models are both desirable products of scientific research and useful as guides to future research" (p. 2). The epistemological functions of models and modelling in generating and communicating scientific knowledge are described extensively. The following compilation shows some of the functions which models and modelling can serve in science according to a range of science philosophers and science education researchers whose work has influenced the present research project (Bailer-Jones, 2002, 2009; Campbell, Oh, Maughn, Kiriazis, & Zuwallack, 2015; Frigg & Hartmann, 2006; Giere, 2004; Leatherdale, 1974; Magnani et al., 1999; Morgan & Morrison, 1999; Odenbaugh, 2005; Oh & Oh, 2011; Passmore et al., 2014; Stachowiak, 1973; van der Valk et al., 2007). The compilation makes no claim to be complete. It rather broadly summarises some aspects which may be linked to the functions models can fruitfully serve in a science classroom.[4]

- Description of phenomena
- Simplification of complex phenomena
- Promotion of imagination
- Communication of ideas
- Development of scientific arguments
- Mediation between theory and phenomenon

[4] A clear assignment of a specific aspect to a philosopher is not feasible for all aspects as the same aspects are being employed overlappingly but through the use of diverse nominations by different philosophers.

- Substitution of theories
- Evaluation of data
- Exploration of systems
- Development of hypotheses
- Formulation and evaluation of predictions
- Knowledge construction
- Theory revision

2.2. Models and Modelling in the Science Classroom

What can one learn with respect to models?

Most science education researchers refer to at least some of the just mentioned points when arguing that models and modelling should be implemented in school science classes (e.g., Acher, Arcà, & Sanmartí, 2007; Campbell & Oh, 2015; Gilbert & Justi, 2016; Grosslight et al., 1991; Namdar & Shen, 2015; Nicolaou & Constantinou, 2014; Oh & Oh, 2011; Schwarz et al., 2009; Upmeier zu Belzen & Krüger, 2010). Passmore et al. (2014) state that "a focus on models and modelling interacts with learning both the conceptual content of particular scientific disciplines and how it may influence students' epistemological views" (p. 1197). Models may be called "effective pedagogical tools" for teaching scientific literacy (Halloun, 2007, p. 653) which help create a more authentic science education environment (Gilbert, 2004; Passmore et al., 2014; Prins, Bulte, & Pilot, 2011). Some researchers also argue that when dealing with models in their class, students gain an understanding about nature of science more broadly (Campbell & Oh, 2015; Gobert et al., 2011). Still, the positive correlation between students' meta-modelling knowledge and their understanding of nature of science as a whole has not been shown empirically.

Passmore et al. (2014) point out that "[t]he practice turn in science education, like the turn in the philosophy of science, has manifested itself in calls for science learning environments to become more authentic to science as actually practiced." (p. 1172).

Standards documents have noted the importance of models in science and have called for an increased role for models in science classrooms. In the U.S. standards documents for science education (NGSS Lead States, 2013; NRC, 2012)[5], 'science practices', which include modelling, are being highlighted as one of the three fundamental learning outcomes alongside 'disciplinary core ideas' and 'cross-cutting concepts'. In Germany, model competence is part of the section 'Erkenntnisgewinnung' which can in parts be compared to 'scientific inquiry' but which explicitly includes aspects of nature of science (KMK, 2005; Mayer, 2007). The Berlin framework curriculum illustrates the aspirations in the aspect of model competence more clearly by distinguishing perspectives which students ought to adopt during their time in school within the three aspects 'using models', 'testing models' and 'changing models' (SenBJF, 2015).[6]

How do we implement models and modelling in the science classroom?

The multiplicity of approaches to implement models and modelling in the science classroom is a consequence of the versatility and the potential of models as both, instruments of science and of pedagogy (Campbell & Oh, 2015; Mittelstraß, 2004; Oh & Oh, 2011) characterised by the perspectives of representation and instrumentation (cf. Mahr, 2011). One way of structuring what students are expected to learn during their time in school is supposed by Hodson (2014) who distinguishes among four major categories

[5] NRC: National Research Council.

[6] SenBJF: Senatsverwaltung für Bildung, Jugend und Familie. The perspectives will be described in more detail in Table 2, as they refer to the theoretical framework of model competence (Krell et al., 2016).

of learning goals: ‘learning science’, ‘learning about science’, ‘doing science’ and ‘addressing socio-scientific issues’ (p. 4). These learning goals can be transferred to the domain of models and modelling and used to categorise approaches which were suggested by science education researchers. Campbell et al. (2015) carried out a review of modelling pedagogies within 81 research articles. They report that 81 % (n = 66) of the articles reviewed contained approaches that aimed at developing conceptual understanding of disciplinary core ideas of science (e.g., Verhoeff, Waarlo, & Boersma, 2008). 30 % (n = 24) of the articles suggested ideas for developing students’ understanding of the nature of models specifically or the nature of science more broadly (e.g., Prins et al., 2011). 10 % (n = 8) of the articles intended to actively engage students in science practice (e.g., Kawasaki, Rupert-Herrenkohl, & Yeary, 2004). Campbell et al. (2015) summarise that “conceptual understanding was the most common pedagogical function identified for modelling, while developing facility and understanding of science practices was identified least often” (p. 159). In terms of Hodson’s (2014) categories, this means that models are more often being used for ‘learning science’ than for ‘learning about science’.

This broad distinction concerning the aims of using models in science classes can be divided further by exposing different pedagogical conceptualisations as to how to translate these aims into learning situations. Campbell et al. (2015) discriminate five conceptualisations for modelling that are based on how scientists employ models in their professional work: (1) exploratory modelling, (2) expressive modelling, (3) experimental modelling, (4) evaluative modelling, and (5) cyclic modelling (cf. Campbell, Oh, & Neilson, 2013; Gilbert & Justi, 2016; Oh & Oh, 2011). The first three were proposed originally by van Joolingen (2004), who describes exploratory modelling as a method where students investigate the properties of a given model by changing parameters and observing the effects of these changes. In expressive modelling, students create new models or they use

existing models to express their ideas and to explain scientific phenomena. In experimental modelling (originally called inquiry modelling), students form hypotheses and predictions from models and test them through experimenting with phenomena. These three basic approaches can be complemented with evaluative modelling which includes the use of multiple representations in order to foster a deeper understanding of phenomena (cf. Ainsworth, 2008) and cyclic modelling, where students are engaged in processes of developing, evaluating, and improving models. Two prominent approaches in cyclic modelling are the so-called GEM cycle (Generation, Evaluation, and Modification of models) by Clement and colleagues (Clement, 1989; Clement & Rea-Ramirez, 2008; Williams & Clement, 2015) and the EIMA proposal (Engage-Investigate-Model-Apply) which focusses on the creation, testing and revision of models by Schwarz and colleagues (Schwarz & Gwekwerere, 2007; Schwarz & White, 2005). The Schwarz group developed a learning progression for scientific modelling which describes increasingly sophisticated ways of understanding a topic by highlighting "aspects of modeling that can be made accessible and meaningful for students and teachers" (Schwarz et al., 2009, p. 633).

Campbell et al. (2015) found that exploratory (43 %) and expressive modelling (54 %) pedagogies, where students engage with models to explain already known scientific phenomena, were those most often leveraged. Pedagogies that focus on students forming hypothesis and conducting investigations on the basis of the model, like experimental modelling (28 %) and cyclic modelling (22 %) where found least often. The diversity of approaches suggested by science education is likely to be a consequence of the diverse functions which models can serve in science and in school.

2.3. Model competence

More and more science education researchers (Schwarz et al., 2009; Upmeier zu Belzen & Krüger, 2010; Vo, Forbes, Zangori, & Schwarz, 2015) call for students to develop a meta-modelling knowledge as a type of nature of science understanding which enables them to reflect about "how models are used, why they are used, and what their strengths and limitations are, in order to appreciate how science works and the dynamic nature of knowledge that science produces" (Schwarz et al., 2009, pp. 634–635).[7] Meta-modelling knowledge can be differentiated from modelling practices due to its explicit reference to students' epistemological awareness about models and the process of modelling (Nicolaou & Constantinou, 2014), while modelling practices focus more on the "how-to" actually perform modelling (cf. Kauertz, Neumann, & Haertig, 2012, p. 713). Within the aspect meta-modelling knowledge, Gilbert and Justi (2016) divide further by distinguishing 'Knowledge about models' (i.e., focus on epistemological and ontological nature of models, reasons why models are constructed and used, how scientific value of models can be assessed) and 'Knowledge about modelling' (i.e., epistemological and ontological grounds on the basis of which models can be constructed, procedures involved in constructing models, procedures involved in evaluating the procedures involved in the construction of models). The combination of both modelling practices and meta-modelling knowledge, complemented with the willingness and motivation to apply model-related capabilities in problematic situations, forms model competence (Gilbert & Justi, 2016; Krell et al., 2016; Nicolaou & Constantinou, 2014; Upmeier zu Belzen & Krüger, 2010). The five

[7] The original idea of meta-modelling knowledge as quoted from Schwarz et al. (2009) is now referred to by the Schwarz group as 'epistemic considerations' which comprise reflections on generality/abstraction, evidence, mechanism, and audience (Vo et al., 2015). In this thesis, I will stick to the original definition as the new one is broader and captures aspects that are not relevant to the present work.

modelling activities (exploratory modelling, expressive modelling, experimental modelling, evaluative modelling, and cyclic modelling; Campbell et al., 2015) may be seen as a pedagogical layer over both the modelling practices and the meta-modelling knowledge.

Table 1. A concept of model competence.[8]

Model Competence	
Modelling Practices	Meta-modelling Knowledge
– Student constructs/ creates a model. – Student compares a model to its original or to other models. – Student uses a model to describe, explain and predict. – Student tests/ evaluates /validates a model. – Student revises a model.	– Student describes the extent to which a model looks like the corresponding original. – Student explains why there are multiple models for one original. – Student describes what purpose a model serves. – Student explains how people can test if a model serves its purpose. – Student names reasons for why a model should be changed.

Model competence has been conceptualised in a number of different theoretical frameworks (Crawford & Cullin, 2005; Grosslight et al., 1991; Justi & Gilbert, 2003; Schwarz, Reiser, Acher, Kenyon, & Fortus, 2012; Treagust, Chittleborough, & Mamiala, 2002; Upmeier zu Belzen & Krüger, 2010). The theoretical framework for the present study is the 'model of model competence', originally published by Upmeier zu Belzen and Krüger (2010) and recently republished by Krell et al. (2016). The 'model of model competence' was develop based on a literature review (Crawford & Cullin, 2005; Grosslight et al., 1991; Justi & Gilbert, 2003; Justi & van Driel, 2005; Treagust et al., 2002, 2004; van der Valk et al., 2007) and theoretical

[8] The table contains aspects proposed by Schwarz et al. (2009), Nicolaou and Constantinou (2014), Gilbert and Justi (2016) as well as by Upmeier zu Belzen and Krüger (2010). A more detailed account of what actions might be included in modelling practices can be found with Gilbert and Justi (2016) as well as Oh and Oh (2011).

reflections about the term ‘model’ (Mahr, 2008, 2009, 2011; Stachowiak, 1973). The framework comprises perspectives which students, teachers and scientists may have towards models and modelling. It can be used to investigate students’ meta-modelling knowledge (Gogolin & Krüger, 2016a; Grünkorn, 2014; Krell, 2013; Patzke, Krüger, & Upmeier zu Belzen, 2015; Terzer, 2012) as well as to evaluate the success of interventions (Fleige et al., 2012; Günther et al., 2016; Orsenne, 2015). Besides, it can give orientation to teachers while designing and realizing modelling-based learning units (cf. Gilbert & Justi, 2016). The ‘model of model competence’ includes the aspects ‘nature of models’, ‘alternative models’, ‘purpose of models’, ‘testing models’ and ‘changing models’ (Table 2). For each of the aspects, the authors propose three levels which reflect ways of understanding the aspect. With reference to Mahr (2008), Upmeier zu Belzen and Krüger (2010) marked the perspective on the model-object itself, the perspective on a ‘model of something’ and on a ‘model for something’. Within each level of each aspect, a model may represent a phenomenon and simultaneously serve as an instrument for a certain purpose (Table 2); a dichotomy that follows from Mahr’s (2008) epistemological relationships that justify something to be conceived of as a model.

According to Upmeier zu Belzen and Krüger (2010), model competence is accredited to students who are able to understand all perspectives in the particular aspect. Upmeier zu Belzen and Krüger (2010) define model competence as follows:

> “Modellkompetenz umfasst die Fähigkeiten, mit Modellen zweckbezogen Erkenntnisse gewinnen zu können und über Modelle mit Bezug auf ihren Zweck urteilen zu können, die Fähigkeiten, über den Prozess der Erkenntnisgewinnung durch Modelle und Modellierungen in der Biologie zu

> reflektieren sowie die Bereitschaft, diese Fähigkeiten in problemhaltigen Situationen anzuwenden."[9] (p. 49)

The words, Upmeier zu Belzen and Krüger (2010) use here to define model competence, underline once again that model competence comprises both modelling practices and meta-modelling knowledge. The definition stresses the notion of a willingness to use the model-related capabilities in problematic situations which is inherent to all competence definitions in the tradition of Weinert (2001).[10]

Table 2. 'Model of model competence' (Krell et al., 2016). Light grey: perspective on model-object, middle grey: perspective on 'model of something', dark grey: perspective on 'model for something' (Mahr, 2008).

	Level I	**Level II**	**Level III**
Nature of models	Model is a replication of the original	Model is an idealised representation of the original	Model is a theoretical reconstruction of the original
Multiple models	Model objects differ	Original allows the creation of different models	Hypotheses about the original differ
Purpose of models	Using the model to describe the original	Using the model to explain something about the original	Using the model to predict something about the original
Testing models	Controlling the model object	Comparing the model to the original	Testing hypotheses about the original with the model
Changing models	Correcting errors in the model object	Revising the model due to new findings about the original	Revising the model due to the falsification of hypotheses about the original with the model

[9] "Model competency includes the capabilities to gain knowledge with models and to judge models with regard to their purpose, the capabilities to reflect upon the process of gaining knowledge through models and modelling in biology, as well as the willingness to apply these capabilities in problematic situations."

[10] A discussion of the terminology of competence, knowledge and skill will follow in chapter 4.2.1. 'Specification of the nature of the construct to be diagnosed' as these considerations are actually part of the validity argument.

How do teachers and students understand models?

Empirical studies indicate that teachers' knowledge of and about models is rather limited (Borrmann, Reinhardt, Krell, & Krüger, 2014; Crawford & Cullin, 2005; Justi & Gilbert, 2003; van Driel & Verloop, 2002). Van Driel and Verloop (1999, 2002) assessed teachers' (*N*=71) understanding of models and modelling using interviews, open-ended tasks and Likert-type scale tasks. The researchers summarise that "the knowledge of the majority of the teachers of models and modelling in science was not very pronounced" (van Driel & Verloop, 1999, p. 1151). A study by Borrmann et al. (2014) with German science teachers (*N*=226) could not replicate the findings of van Driel and Verloop (1999). Still, the study by Borrmann et al. (2014) showed that the questioned teachers mainly attributed purposes of illustration and explanation to models ('models of something') rather than hypothesis testing ('models for something'). The authors similarly conclude that the teachers' understanding of models is not elaborate. Smit and Finegold (1995) investigated perceptions held by final-year preservice physics teachers about models and stated that

> "55 % of the respondents saw a model as a copy/replica/imitation or reflection of something that exists in nature. The model is seen as identical or nearly identical to the real thing. This view of a model is a very restricted one and can have serious implications for the teaching of the subject." (p. 630).

Crawford and Cullin (2004) likewise report that the teachers (*N*=16) who took part in their model-based instructional module did not appear to achieve full understanding of scientific modelling (Crawford & Cullin, 2004). Although the researchers detected positive changes, preservice teachers held on to some scientifically uninformed views (Crawford & Cullin, 2005). Justi and Gilbert (2002, 2003) collected data pertaining to teachers' meta-modelling knowledge and metacognitive knowledge about

modelling by conducting interviews with science teachers (*N*=39). The researchers conclude that the teachers do not hold coherent ontological and epistemological views of models and modelling (Justi & Gilbert, 2003). They report that the teachers see the purpose of modelling in constructing a model which is almost identical to an already known phenomenon. Justi and Gilbert (2002) found that teachers, despite declaring themselves in favour of focusing on the nature of models and modelling, "placed great importance on learning the content of specific scientific/historical models" (p. 1276) and consequently use models to 'learn science'. The authors also point out that the teachers' ideas about the use of models in the science classroom are influenced by what they believe their students to think of models. Unfortunately, teachers did not know or simply guessed what their students think on this matter (Justi & Gilbert, 2002).

Students' meta-modelling knowledge has been subject to a great number of studies using different theoretical backgrounds and diverse methodological approaches, ranging from interviews to closed-ended tasks (Al-Balushi, 2011; Chittleborough, Treagust, Mamiala, & Mocerino, 2005; Gobert et al., 2011; Grosslight et al., 1991; Grünkorn, 2014; Krell, 2013; Krell, Upmeier zu Belzen, & Krüger, 2012; Lee, Chang, & Wu, 2015; Patzke et al., 2015; Pluta, Chinn, & Duncan, 2011; Schwarz et al., 2009; Schwarz & White, 2005; Sins, Savelsbergh, van Joolingen, & van Hout-Wolters, 2009; Terzer, 2012; Treagust et al., 2002, 2004; Trier et al., 2014). Due to the abundance of studies and approaches to empirically detecting students' meta-modelling knowledge, I will describe merely those studies which most prominently contributed to the development and to the evaluation of the diagnostic instrument of the present research project.[11] In the following continuous text, I will outline the chronological and content-related connections

[11] Further reviews are assembled by Campbell et al. (2015); Gilbert and Justi (2016); Krell et al. (2016); Namdar and Shen (2015); Nicolaou and Constantinou (2014); Oh and Oh (2011).

between the studies. The main results of the studies are summarised in Attachment 1 and will be referred to throughout this work when justifying methodical decisions or discussing findings.

The exploratory interview study with seventh and eleventh graders by Grosslight et al. (1991) is considered fundamental to research on students' meta-modelling knowledge as it is the first to describe distinct aspects and global levels of understanding of models and modelling. The levels were developed based on Carey, Evans, Honda, Jay, and Unger's (1989) analysis of students' epistemological understanding of science ranging from a naive-realist (level I) via a relativist (level II) to a constructivist epistemology of science (level III). In Grosslight et al.'s (1991) study, most of the students ranged within levels I and II, while none of the interviewed students reached a level III understanding. The global levels include perspectives on six different aspects of models and modelling and thereby "summarise the kind of understanding and conceptions of models that emerged from the interview as a whole" (Grosslight et al., 1991, p. 817). These global levels of understanding were later criticised as empirical studies provided more support for the definition of aspect-dependent levels (Crawford & Cullin, 2005; Justi & Gilbert, 2003; Krell, Upmeier zu Belzen, & Krüger, 2014c).

Treagust and colleagues (Chittleborough et al., 2005; Treagust et al., 2002, 2004) investigated students' meta-modelling knowledge using primarily closed-ended tasks combined with some qualitative methods as measures of control and specification. They created the 'Students' Understanding of Models in Science (SUMS)' Likert-type questionnaire, the 'Molecular Representations (MR)' Likert-type questionnaire and the 'My Views of Models and Modelling in Science (VOMMS)' forced choice questionnaire which evolved from Aikenhead and Ryan's (1992) questions on 'Views of Science-Technology-Society' (VOSTS). Summarising, the studies of Treagust and colleagues (Chittleborough et al., 2005; Treagust et al., 2002, 2004) paint a more detailed and more positive picture of students' meta-

modelling knowledge than the previous findings by Grosslight et al. (1991). Chittleborough et al. (2005) provide further evidence for students' meta-modelling knowledge to improve across years in school. Another achievement of the work of Treagust and colleagues is to highlight the necessity of respecting the type of model used as a context for the assessment, whether it may be more or less abstract (Treagust et al., 2002) or whether it belongs to the domain of science or school (Treagust et al., 2004). In an interview study, Trier et al. (2014) showed similarly that students distinguish between school and science contexts. Al-Balushi (2011) picked up on the claim for differentiation across contexts and across grades and asked students from grades 9 to 11 as well as preservice science teachers to rate the credibility of a range of abstract to concrete models. Al-Balushi (2011) summarises that the students in higher grade levels rated fewer models as certain and more as imaginary but at the same time, the students answered inconsistently across different contexts. In a study by Krell et al. (2012) focusing on the purpose of biological models, students' meta-modelling knowledge equally varied across task contexts.

Being based on the 'model of model competence' by Upmeier zu Belzen and Krüger (2010), the forced choice tasks by Krell (2013) and the open-ended tasks by Grünkorn (2014) can be considered as the direct predecessors of the present study. The tasks of both projects were used to generate empirical evidence to validate the perspectives within the levels of the 'model of model competence'. Next to assessing students' meta-modelling knowledge within the five aspects proposed by the theoretical framework (Table 2), Krell's (2013) investigations concerning the dimensionality of the framework underlined the need for a more differentiated aspect-dependent assessment and promotion of students' meta-modelling knowledge. Grünkorn (2014), meanwhile, provided evidence that the students' perspectives within the aspects 'nature of models' and 'purpose of models' were sufficiently described by the 'model

of model competence', meaning that no 'initial level' of understanding had to be added, whereas in the aspects 'multiple models', 'testing models' and 'changing models', additional perspectives on 'initial levels' should be considered.

2.4. Test Development

All of the just mentioned approaches to assess teachers' and students' meta-modelling knowledge are based on content-related and methodical considerations concerning test development and evaluation. Unfortunately, often in educational assessment, information concerning the soundness of the proposed score interpretation (e.g., details about the construction process, empirical evidence for validity) are limited (Karabenick et al., 2007; Leighton, 2004). During the development of the instrument for the present research project, we aimed at providing as much evidence for sound score interpretation as possible within the time we had and with focus on claims that we considered inherent in the interpretation and use of the test (Kane, 2015). In the following passage, I will point out possibilities to gather evidence for validity during test development.

Considering solely content specifications to describe an objective (e.g., behaviour, understanding, development) leads to questions about proposed test score interpretations along the lines of reliability, generalizability and validity (Messick, 1995). On the other side, Anastasi (1976) remarks for the field of psychology that "those psychologists specializing in psychometrics have been devoting more and more of their efforts to refining techniques of test construction, while losing sight of the behaviour they set out to measure" (p. 297).[12] In his 'Framework for Developing Cognitively

[12] A comprehensive argumentation concerning the balance between content-related and methodical considerations concerning test development with reference to historical

Diagnostic Assessments', Nichols (1994) picks up on this antagonism and advertises five steps for psychology-driven test development for education which comprise (1) 'Substantive theory construction', (2) 'Design selection', (3) 'Test administration', (4) 'Response scoring' and (5) 'Design revision'. The purpose of all of these steps is to gather evidence for validity by examining whether the interpretations of test scores support the model or theory on which they are based. Nowadays, establishing validity is considered the central element of test development (AERA et al., 2014).

Validity[13]

Nichols (1994) highlights with reference to validity that, "as with any scientific theory, the theory used in development is never proven; rather, evidence is gradually accumulated that supports or challenges the theory" (p. 587). This notion of validity is also the basis of the Standards for Educational and Psychological Testing (AERA et al., 2014):

> "Validity refers to the degree to which evidence and theory support the interpretations of test scores for proposed uses of tests. Validity is, therefore, the most fundamental consideration in developing tests and evaluating tests. The process of validation involves accumulating relevant evidence to provide a sound scientific basis for the proposed score interpretations. It is the interpretations of test scores for proposed uses that are evaluated, not the test itself." (p. 11)

For modern test development, it is crucial to denote that the current definition of test validity which focusses on the interpretation of scores

development and future implications can be found in the book 'Cognitive Diagnostic Assessment for Education: Theory and Applications' (Leighton and Gierl, 2007a).

[13] In the modern understanding of validity (AERA et al., 2014), the latter includes reliability and fairness in testing as aspects of evidence based on internal structure but on account of the scope of these points alone, on account of their role in the classical tripartism of objectivity, reliability and validity, as well as on account of the transparency of the theoretical information, reliability and fairness in testing will be presented separately.

rather than on the test itself differs from the traditional idea where validity defined the extent to which a test measures what it is intended to measure (cf. Schmiemann & Lücken, 2014). Borsboom and Markus (2013) illustrate why this difference is important by the metaphor of an alarm clock. According to the traditional notion, "an alarm clock is reliable if it sounds consistently at some specific time and valid if it sounds consistently at the intended time" (p. 110). According to the current notion of validity, the "conclusion that the time has come to get up follows appropriately from the sounding of the alarm if the inference is valid" (Borsboom & Markus, 2013, p. 110). As it is the conclusion which we are interested in, we should evaluate the soundness of the conclusion rather than the compliance to the intention of the measurement.

Both Nichols (1994) and the Standards for Educational and Psychological Testing (AERA et al., 2014) point out the need for an accumulation of diverse pieces of evidence for the soundness of proposed score interpretations. The Standards name four distinct sources of evidence: 'Evidence Based on Test Content', 'Evidence Based on Response Processes', 'Evidence Based on Relations to Other Variables' and 'Evidence Based on Internal Structure'. All of these evidences are to be weighted with respect to the overarching aspect of 'Consequences of testing' (AERA et al., 2014) and evaluated against the background of the hypothesis of the proposed interpretation of scores. Each of these sources accumulates evidence to challenge the hypothesis.

According to the Standards for Educational and Psychological Testing (AERA et al., 2014), 'Evidence Based on Test Content' can be obtained from a logical or empirical analysis of the relationship between the content of a test and the construct it is intended to measure. A careful expert review of the construct and test content domain may point to potential sources of irrelevant difficulty (or easiness) that require further investigation. Messick (1995) highlights two major threats to validity: (1) 'construct

underrepresentation' where the assessment is too narrow and fails to include important dimensions or facets of the construct, and (2) 'construct-irrelevant variance' where "the assessment is too broad, containing excess reliable variance associated with other distinct constructs as well as method variance such as response sets or guessing propensities that affects responses in a manner irrelevant to the interpreted construct" (p. 742). According to Messick (1995), both of these threats are operative in all assessments and it is the mission of the test developer to examine and control for these threats. Floyd, Phaneuf, and Wilczynski (2005) provide a list of methods for examining aspects of validity including 'Evidence Based on Test Content' and name for example: "development of items based on literature review, review of other instruments, evaluation of item clarity, readability analysis of items, readability analysis of written informant instructions […]" (p. 8).

'Evidence Based on Response Processes' concerns the fit between the construct and the detailed nature of the performance or response actually engaged in by test takers (AERA et al., 2014). Messick (1995) calls for querying respondents about their solution processes (student interviews) or asking them to think aloud while responding to exercises during field trials.[14] The importance of integrating the responses of the target sample group into the process of test development is equally pointed out by Cronbach and Meehl (1955) who argue that the description of expertise in a certain content domain (model of domain mastery; cf. Leighton, 2004) and its translation into a set of tasks (model of test specifications; cf. Leighton, 2004) function as hypotheses until students' actual thoughts and responses to a final set of tasks have been investigated.[15]

[14] An example of how these two suggestions may be integrated can be found with Adams and Wieman (2011) who propose a development framework for instruments to measure learning of expert-like thinking.

[15] A comprehensive summary of the concept and use of cognitive models of task performance including models of domain mastery and test specification has been published by Leighton (2004).

‘Evidence Based on Relations to Other Variables’ relies on the comparison of the test under investigation with external variables. Raykov and Marcoulides (2011) point out that “Constructs cannot be defined only in terms of operational definitions but also must demonstrate relationships (or lack thereof) with other constructs and observable phenomena” (p. 8). Evidence based on relations to other variables includes two major categories of validity evidence which were historically referred to as construct validity and criterion validity. The former contributes analyses of convergent and discriminant validity (e.g., mixed and multi methods research; Morse, 2003).

> “Relationships between test scores and other measures intended to assess the same or similar constructs provide convergent evidence, whereas relationships between test scores and measures purportedly of different constructs provide discriminant evidence” (AERA et al., 2014, p. 17).

Evidence may also stem from external variables which are mostly predictive in nature (e.g., job success; McCoach, Gable, & Madura, 2013). Known-Groups comparisons are listed by Floyd et al. (2005) as a measure to demonstrate whether a test can discriminate between groups known to differ on the variable of interest.

‘Evidence Based on Internal Structure’ can be obtained by analysing whether theoretically expected patterns of relationships among tasks and/or test components can be identified. Evidence is gathered when the patterns confirm the construct on which the proposed test score interpretations are based (AERA et al., 2014; Messick, 1995). Typically, internal structure refers to aspects of dimensionality, measurement invariance and reliability (Rios & Wells, 2014). In recent years, possibly as a side effect of the ‘No Child Left Behind Act’ in the U.S., the aspect of fairness in testing for subgroups gained increasing importance in test development as part of

evidence based on internal structure (AERA et al., 2014; Kunnan, 2007; Sadovnik, O'Day, Bohrnstedt, & Borman, 2013).

Reliability

The idea of reliability is said to have been introduced by Spearman (1904) who noticed that accidental errors of measurement meant that relations between observed scores were less than the true relations between the underlying variables (Field, 2013). According to Field (2013), reliability measures indicate whether an instrument can be interpreted consistently across different situations, so to say the ability of the measure to produce the same result under the same conditions.

The Standards for Educational and Psychological Testing (AERA et al., 2014) prefer the term 'reliability/precision' to signify the more general notion of consistency of the scores across instances of the testing procedure. They point out that test developers need to have some indication of the reliability of the scoring procedure which is coupled to the developed test and which enables the scorer to evaluate the measured behaviour. The Standards (AERA et al., 2014) refer to the consequences of testing when interpreting measures of reliability:

> "The need for precision increases as the consequences of decisions and interpretations grow in importance. If a test score leads to a decision that is not easily reversed, […] a higher degree of reliability/precision is warranted. If a decision can and will be corroborated by information from other sources or if an erroneous initial decision can be easily corrected, scores with more modest reliability/precision may suffice." (p. 33)

Similar to the decision making for validity, according to the Standards for Educational and Psychological Testing (AERA et al., 2014), analyses of reliability/precision depend on the variability allowed in the testing procedure (e.g., contexts, formats) and the proposed interpretation of test

scores. For example, under the assumption that the assessed construct does not vary over occasions, the variability over occasions is a potential source of measurement error. Cronbach and Shavelson (2004) criticise that "Coefficients are a crude device that do not bring to the surface many subtleties implied by variance components." (p. 394). Therefore, it becomes all the more important to investigate the theoretical construct to be measured as precisely as possible in order to be able to generate sound assumptions for the evaluation of a test instrument. According to the Standards for Educational and Psychological Testing (AERA et al., 2014), changes in scores from one occasion to another are not regarded as error if they result, in part, from changes in the construct being measured.

Reliability is usually being estimated and interpreted in the course of test development by means of coefficients, including reliability coefficients, generalizability coefficients and IRT information functions (Raykov & Marcoulides, 2011). Classical test theory recognises primarily three coefficients: (1) 'alternate-form coefficients' where parallel forms of the same test are administered, (2) 'test-retest coefficients' where the same form of the test is administered on separate occasions and (3) 'internal-consistency coefficients' which are based on the correlation among individual items within a test (AERA et al., 2014; Raykov & Marcoulides, 2011). All of them are considered simultaneously in generalizability theory to create a standard error of measurement (Webb, Shavelson, & Haertel, 2007). For example, standard errors can be estimated for different observed scores based on test information functions from IRT models (Adams, 2005; AERA et al., 2014).

Fairness in Testing

Just like reliability, fairness in testing is part of validity, more precisely, a fair test provides evidence for validity based on internal structure. The

Standards for Educational and Psychological Testing (AERA et al., 2014) describe fairness in testing as follows:

> "A test that is fair within the meaning of the Standards reflects the same construct(s) for all test takers, and scores from it have the same meaning for all individuals in the intended population; a fair test does not advantage or disadvantage some individuals because of characteristics irrelevant to the intended construct." (p. 50)

When the access to the tasks is impeded by reasons unrelated to the construct to be measured, this limits the validity of score interpretations for intended uses for some individuals or subgroups. Whereas traditionally, the quality criterion of objectivity asked for standardised test administration conditions as well as clear scoring procedures, modern test development makes use of the approach of 'Universal Design' which seeks to maximise accessibility to the test for all intended users by clearly defining the construct to be measured, the purpose for which scores will be used, the inferences that will be drawn from the scores and the characteristics of the intended test population with special focus on possible subgroups (AERA et al., 2014; Raykov & Marcoulides, 2011). Pragmatically, fairness may be seen as a lack in measurement bias where "characteristics of the test itself that are not related to the construct being measured, or the manner in which the test is used, may sometimes result in different meanings for scores earned by members of different identifiable subgroups." (AERA et al., 2014). According to the Standards for Educational and Psychological Testing (AERA et al., 2014), threats to fair interpretations of test scores derive from aspects of the test that produce construct-irrelevant variance (Messick, 1995), such as (1) inappropriate sampling of test content (e.g., motivational differences across tasks), (2) disparities in test context (e.g., lack of clarity in test instructions, differences in item complexities, scoring criteria that favour one group over another), (3) variety of test responses

(e.g., social desirability) and (4) opportunity to learn (e.g., exposure of the test taker to instruction or experiences with regard to the construct to be measured). The Standards for Educational and Psychological Testing (AERA et al., 2014) highlight that it is often not clear whether the score differences are a consequence of real differences between groups within the assessed construct or if they are due to some source of bias. Most often, so the Standards (AERA et al., 2014), it may be some combination of real differences and bias.

3. Research Questions

The relevance of models in the scientific community and the demand for their integration into school curricula have lead researchers to investigate into students' meta-modelling knowledge. It is reported that students have difficulties acknowledging the role models play as instruments in science (chapter 2.3.). The 'model of model competence' can serve as a basis for the development of diagnostic instruments which in return may help teachers plan individualised interventions. During test development, evidence for the validity of the future score interpretations needs to be collected, weighted and integrated in order to form a unitary validity argument (chapter 2.4.).
Science teachers who wish to improve their teaching about models and modelling need to know what their students think on the matter (e.g., Oh & Oh, 2011; Campbell et al., 2016; cf. Duit, Gropengießer, Kattmann, Komorek, & Parchmann, 2012; Hartig et al., 2008). With reference to the 'Assessment of Competencies in Educational Contexts', Hartig et al. (2008) stress the following:

> "The success of many educational decisions and interventions hinges on accurate assessments of learners' baseline competencies and learning outcomes. […] Assessment to promote individual learning can be regarded as formative evaluation on an individual level. It should allow precise conclusions to be drawn about individual learning processes and learners' strength and weaknesses with respect to specific curricular units. These conclusions can help to support individual instruction and learning and ideally offer considerable potential to enhance teaching." (p. 15)

An instrument that provides teachers with direct feedback on their students' meta-modelling knowledge of the nature and the purpose of models could help determining specific starting points for effective interventions (Hartig et al., 2008). Existing instruments cannot provide this service sufficiently well (Kauertz et al., 2012).

The aim of the present research project is to develop a test instrument for students' meta-modelling knowledge of the nature and the purpose of models on the theoretical basis of the 'model of model competence' (Krell et al., 2016) and on the empirical basis of students' responses to open-ended diagnostic tasks (Grünkorn, 2014). The main theoretical pillars for such an endeavour are reflected in the previous chapter 'Theoretical Frame and Scope' which concerns both content-related questions about models and modelling as well as questions concerning test development. The following research questions and hypotheses are equally twofold, concerning both, test development and diagnosis. In the present chapter, I will outline global research questions and hypotheses for the whole of the research project. The entirety of the findings ought to be discussed with reference to these global questions in chapter 5. In the course of this thesis, there will be a number of secondary research questions and hypotheses that refer specifically to the studies performed during the research project and for which the discussion from associated articles will be summarised.

3.1. Test Development

Crocker and Algina (1986) state that "in test construction, a general goal is to arrive at a test of minimum length that will yield scores with the necessary degree of reliability and validity for the intended uses" (p. 311). The Standards for Educational and Psychological Testing (AERA et al., 2014) give instructions as to how to arrive at such a test.

The tasks which were developed in this project based on the instructions by the Standards are to be put to the test as to whether they sustain the demands proposed by validity. The first global research question (RQ_{GI}) subsequently concerns the soundness of the proposed score interpretations:

RQ_{GI} In what way do evidence and theory support the interpretation of the test scores for the intended use of diagnosing students' meta-modelling knowledge?

The investigation of RQ_{GI} involves considering multiple sources of evidence for validity. The following global hypotheses (H_G) operationalise validity of the score interpretations as the superordinate hypothesis which is to be challenged by investigations.

Evidence for validity based on test content

H_{GI} 1 The test content adequately represents the content domain.

Evidence for validity based on response processes

H_{GI} 2 The expected performances measured by the test are actually performed by students when taking the test.

Evidence for validity based on relations to other variables

H_{GI} 3 Theoretically expected patterns of relationships of the test under investigation with external variables can be identified in students' test scores.

Evidence for validity based on internal structure

H_{GI} 4 Theoretically expected patterns of relationships among test components can be identified in students' test scores.

3.2. Diagnosis of Students' Meta-modelling Knowledge

Given the premise that sufficient evidence for a valid proposed score interpretation for the tasks was accumulated in the test development process (RQ_{GI}), the final instrument is to be used to diagnose students' meta-modelling knowledge in the aspects 'nature of models' and 'purpose of models' with reference to the theoretical basis of the 'model of model competence' (Krell et al., 2016). In order to utilise the diagnostic instrument to supply teachers with content-rich information about students' meta-modelling knowledge, we raise the second global research question:

RQ_{GII} How frequently are the three levels of understanding within the aspects 'nature of models' and 'purpose of models' represented among students?

Findings for students' meta-modelling knowledge in the aspects 'nature of models' and 'purpose of models' are diverse across empirical studies and the two aspects ought to be regarded separately. The studies methodically closest to the present research project are those that employed closed-ended tasks (Chittleborough et al., 2005; Krell, 2013; Treagust et al., 2002, 2004). These studies showed that their investigated students had a more elaborate understanding than expected from many other previous studies including the ones of Grosslight et al. (1991) and Harrison and Treagust (1996).

For the aspect 'nature of models', Treagust et al. (2004) as well as Chittleborough et al. (2005) report that more than 80 % of their eleventh grade students see models as representations of ideas of how things work (Attachment 1), a perspective which could be aligned with a level III perspective in the 'model of model competence'.

$H_{GII\ Nature}$ The majority of students show a meta-modelling knowledge of the nature of models on level III.

For the aspect ‘purpose of models’, the works by Treagust et al. (2004) as well as Chittleborough et al. (2005) are less bountiful because the direct alignment of the VOMMS instrument with the conception of the aspect ‘purpose of models’ from the ‘model of model competence’ is not feasible. Therefore the results presented by Krell (2013), who’s methodology is very close to the present project, although divergent in the grade of the sample, will be used to derive a hypothesis concerning students’ meta-modelling knowledge of the purpose of models. Krell (2013) found that the majority (37 %) of the students in grades seven to ten chose an answer on level III in the aspect ‘purpose of models’ indicating that models can be used for predictions (Attachment 1).

$H_{GII\ Purpose}$ The majority of students show a meta-modelling knowledge of the purpose of models on level III.

4. Gathering Evidence during Test Development

Once aim and scope are laid out, it is necessary to delineate fundamental decisions and findings that occurred during the development of the diagnostic instrument. Unveiling and giving reason for certain decisions along the process of research is a central quality criterion of test development as it forms the basis for later interpretation of the results and gives other researchers the opportunity to judge upon the validity of the proposed score interpretations (AERA et al., 2014; Döring & Bortz, 2016; Raykov & Marcoulides, 2011).

In chapter 4.1, I will provide a chronological tabular overview of the studies conducted during test development. Instead of proceeding chronologically, the chapters 4.2 to 4.5 refer to findings with regard to the four great sources of evidence for validity named by the Standards for Educational and Psychological Testing (AERA et al., 2014). The focus of this report is on the contribution of these different evidences to the overall construction process and validity argument rather than on the protocol-like record keeping of the test development process. In some chapters, studies at different stages of the development process will be included. So to increase transparency, for every study, the version of answer options that was used will be indicated. Most of the findings have also been published in research articles and are summarised here with reference to the articles which can be found in chapter 11. In chapter 5, all evidence will be integrated and discussed so to make a final evaluation of the validity of the proposed future score interpretations. This also involves a discussion about the consequences of testing as requested by the Standards for Educational and Psychological Testing (AERA et al., 2014).

4.1. Studies during Test Development

Next to theoretical considerations, a number of studies have helped testing the hypothesis of validity empirically. The studies are summarised chronologically in Table 3. The table protocols the study design, indicates which version of the tasks was used in each study and what kind of validity evidence was to be gathered. The single studies will be summarised in the following chapters with an indication of the articles that they have been reported in.

Table 3. Studies that helped optimising the instrument and gathering evidence for validity.

Study	Study design	Tasks	*N*	Consequences for test development	Evidence for validity …	Publication
I	Expert judgement study with answer options for each of 6 contexts in the aspect ‘nature of models’	$N1_9$	11	Reformulation of answer options; Change of 1 context; Addition of 2 contexts	… based on test content	-
II	Pilot study with forced choice tasks for 8 contexts in the aspect ‘nature of models’	$N2_{6\text{-}9}$	471	Reduction of answer options per context from nine to six	… based on test content	Article 1
III	Monotrait-multimethod study with forced choice tasks, open-ended diagnostic tasks and interviews	$N3_6$	448	Forced choice tasks only suitable for grades 11 and 12	… based on relations to other variables	Article 2
IV	Think aloud study with forced choice tasks in the aspect ‘nature of models’	$N3_6$	30	Reduction of answer options per context from six to three; Dismissal of two contexts	… based on response processes	
V	Expert judgement study with single answer options for each of 8 contexts in the aspect ‘purpose of models’	$P1_6$	9	Reformulation of answer options	… based on test content	-
VI	Mixed method study with single answer options for each of 8 contexts in the aspect ‘purpose of models’ on Likert-scales; subsequent student interviews	$P2_6$	275	Reduction of answer options per context from six to three; Dismissal of two contexts	… based on test content and on response processes	Article 3
VII	Mixed method study with final diagnostic instrument with forced choice tasks in the aspects ‘nature of models’ and ‘purpose of models’; open-ended justification tasks	$N3_3$ / $P3_3$	382	Basis for Diagnostic Algorithm (Regression analysis)	… based on internal structure	Article 4
VIII	Intervention study; final diagnostic instrument with forced choice tasks in the aspects ‘nature of models’ and ‘purpose of models’	$N3_3$ / $P3_3$	65		… based on relations to other variables	Article 6

Note: The column ‘Tasks’ refers to the version of answer options used in the study. The different versions of the answer options for each aspect are attached in chapter 10.5. The index following the task version indicates the amount of answer options per context.

4.2. Evidence Based on Test Content

Nichols' (1994) five steps for psychology-driven test development for education comprise 'Substantive theory construction' and 'Design selection'. In terms of validity, these steps basically cover what is called 'Evidence Based on Test Content' by the Standards for Educational and Psychological Testing (AERA et al., 2014). They require of test developers to clearly describe the construct that the test is intended to assess and to analyse the relationship between the content of a test and the construct.
Such considerations built the starting point for task development in this research project. In the present chapter, I will describe the fundamental decisions taken before and in the first stage of task development. This mainly involves theory-based considerations about the nature of the construct that the diagnostic instrument is intended to diagnose and about the task format but also empirical studies with panels of experts.

4.2.1. Specification of the nature of the construct to be diagnosed

The theoretical basis of the diagnostic instrument is the 'model of model competence' (Krell et al., 2016). The concept of 'competence' has been widely used to describe capabilities to master a certain domain (Kauertz et al., 2012). Upmeier zu Belzen and Krüger (2010) refer to Weinert (2001) who defines competence as the sum of available or learnable abilities and skills together with willingness to solve upcoming problems and to act responsibly and critically concerning the solution. Other definitions facilitate the assessment of the construct of competence by limiting the view to a cognitive perspective (Hartig et al., 2008; Hartig & Klieme, 2007; Kauertz et al., 2012; Koeppen, Hartig, Klieme, & Leutner, 2008). Hartig et

al. (2008) advocate the explicit separation of cognitive and motivational aspects by arguing that motivation may change over time while competences are seen to be more stable. This separation is also recommended by Weinert (2001) for empirical studies and subsequently adopted for the present study.

Considering that model competence is made up of ‘modelling practices’ and ‘meta-modelling knowledge’ (

Table 1), the instrument of the present project is set to diagnose the latter. Students are not being asked to manually construct models or to work with existing models but to reflect about models on a meta-level. In order to avoid confusion, I may point out that this project refers to the aspects of meta-modelling knowledge as proposed by Upmeier zu Belzen and Krüger (2010) and not as proposed by Schwarz et al. (2009) who structure meta-modelling knowledge differently. The students in the present study are specifically asked to reflect about the aspects ‘nature of models’ and ‘purpose of models’ described as two of five aspects in the ‘model of model competence’. The students have to state to what extent a model corresponds to an original (‘nature of models’) and which purpose a model may serve (purpose of models). The decision to choose the two mentioned aspects ‘nature of models’ and ‘purpose of models’ out of the five aspects proposed by Upmeier zu Belzen and Krüger (2010) is based on several considerations. The previous work by Terzer (2012) suggests the aspect ‘nature of models’ to be a global factor on which all other aspects are based as it was highly correlated with all of the other aspects. As for promotion, Terzer (2012) argues to explicate, at first, the nature of models while keeping in mind the purpose of models. The aspect ‘nature of models’ focusses exclusively on students’ perspectives on the relationship between model and original which in return may influence students’ perspectives in other aspects. Grosslight et al. (1991) as well argue that a naïve understanding of the nature of models is likely to influence the perspectives

on models in general to be rather medial. Out of the five aspects, specifying the nature of the relationship between a model and the world represents the question most intensively debated about in science philosophy (Passmore et al., 2014). Tasks for the aspect ‘purpose of models’ were added because we observed in our studies for the aspect ‘nature of models’ that the vast majority of students named the proposed purpose of the models when reflecting about their nature. This observation had equally been made by Grünkorn (2014) in previous studies with open-ended tasks. Mahr and Wendler (2009) strikingly logically argue the dependence between both aspects:

> „Nach dem im Modell des Modellseins explizierten Modellbegriff ist ein Gegenstand ein Modell, wenn er als Modell aufgefasst wird. Da grundsätzlich jeder Gegenstand als Modell aufgefasst werden kann, ist kein Gegenstand schon für sich ein Modell, sondern erst durch das Urteil eines Subjekts. Mit diesem Urteil wird ein Gegenstand in einer konkreten Situation dadurch zu einem Modell, dass ihn das urteilende Subjekt in einen konkreten oder gedachten Zusammenhang (Kontext) stellt, der ihn zum Modell macht.“ (p. 1)[16]

Addressing these two aspects allowed us to focus on both, students’ ‘Knowledge about models’ and their ‘Knowledge about modelling’ (Gilbert & Justi, 2016; Upmeier zu Belzen & Krüger, 2010). The combination of the aspects ‘nature of models’ and ‘purpose of models’ respects the way models are employed by scientists (Bailer-Jones, 2002; Boumans, 1999; Odenbaugh, 2005; van der Valk et al., 2007) and appreciated by philosophers within the semantic view (Bailer-Jones, 2009; Giere, 2001; Magnani et al., 1999; Mahr, 2008).

[16] According to the epistemic pattern of model being, an object becomes a model when being conceived of as a model. Since basically every object can be conceived of as a model, no object is in itself a model, but through a subjects judgement. The judgement transforms an object into a model by a subject putting it in a concrete or hypothetical context.

Despite all considerations concerning the nature of the construct to be assessed, it seems important to point out that when putting students to the test (even when merely focussing on the cognitive aspects of competences) one should separate competence and performance as competences cannot be tested as such (Messick, 1984; Schecker, 2012; Sophian, 1997). Messick (1984) puts it in the following words:

> „Competence refers to what a person knows and can do under ideal circumstances, whereas performance refers to what is actually done under existing circumstances. Competence embraces the structure of knowledge and abilities, whereas performance subsumes as well the processes of accessing and utilizing those structures and a host of affective, motivational, attentional and stylistic factors that influence the ultimate responses.“ (p. 227)

Students might fail answering certain tasks but not because they lack the meta-modelling knowledge the task is intended to assess, but because of aspects of the testing situation (Sophian, 1997, p. 281). This means that we need to be careful when interpreting the performance of students in order to draw conclusion about their competence or about aspects of that competence. Ultimately, it leads us right back at the importance of establishing as many arguments for a valid score interpretation as possible.

4.2.2. Choice of task format

When deciding for a task format for a diagnostic instrument, we first had to weigh our options of assessment methods. Nowadays, many researchers apply mixed method designs for test development or for intervention studies (Creswell & Plano Clark, 2011). Nevertheless, the final diagnostic instrument most likely contains either closed-ended tasks (Krell, 2013;

Terzer, 2012; Treagust et al., 2002), open-ended tasks (Grünkorn, 2014), hands-on tasks (Orsenne, 2015), interviews (Grosslight et al., 1991; Trier et al., 2014) or a combination (Treagust et al., 2004). For this study, the format of closed-ended tasks was chosen due to its advantages over open-ended tasks concerning efficient data collection and analysis (cf. Grünkorn, 2014). Considerations concerning the classical paradigm debate including strengths and weaknesses of qualitative and quantitative methods have been summarised by Choy (2014), Johnson and Onwuegbuzie (2004) as well as Yauch and Steudel (2003).

Within quantitative approaches, there is a great variety of closed-ended task formats to choose from (e.g., rating tasks, single/multiple choice tasks, forced choice tasks, matching tasks, true-false task, rearrangement tasks; Moosbrugger & Kelava, 2012). Some of the aspects considered during test development with regard to the three task formats of rating tasks, multiple choice tasks and forced choice tasks are summarised in Table 4. Only these three tasks formats were chosen for comparison because they had been used in previous studies investigating students' understanding of models so the results of studies could be more easily compared using one of these formats.

Table 4. Comparison of closed-ended task formats.[17]

Aspect	**Rating task**	**Multiple choice task**	**Forced choice task**
Effort for working memory	low (every answer option is rated individually)	high (all answer options need to be compared)	high (all answer options need to be compared)
Nature of answer options	preconceived concepts; either right and wrong or standings along a construct continuum	preconceived concepts; right and wrong	preconceived concepts; standings along a construct continuum

[17] Table 4 shows a compilation of aspects presented by Bennett (1991); Böckenholt (2004); Chaoui (2011); Döring and Bortz (2016); Famularo (2007); Gorin (2007); Krell and Krüger (2011); Martinez (1999); McCloy, Heggestad, and Reeve (2005); Moosbrugger and Kelava (2012).

Exhaustivity	Not necessary	Necessary	Not necessary
Clarity of interpretation	Tied judgements possible	No tied judgements possible	No tied judgements possible
	Does not force student to take a stand, as multiple answer options could be rated the same	Forces student to take a stand	Forces student to take a stand
	Prone for answer tendencies (extreme, middle)	Not prone for answer tendencies	Not prone for answer tendencies

Note: Advantages of a certain task format (for our purpose) ahead of one or both of the other formats are marked grey.

The choice of task format depends on the purpose of the diagnostic instrument. Our purpose was to be able to directly interpret the answer of a student as a level from the 'model of model competence' (Krell et al., 2016). We decided against Likert-type rating scales because with this task format, it would be possible and intelligible for students to rate answer options representing different levels equally and thereby the tasks would not force students to take a stand (Böckenholt, 2004). Conceptually, this would not be a problem because all three answer options are appropriate but it would make the clear interpretation of the score and the recommendation for promotions difficult. Multiple choice tasks avoid the problem of tied judgements but they request for the exhaustivity of answer options meaning that either all answers have to be wrong (distractors) except for one (attractor) or all possible perspectives on a matter need to be given ('ordered multiple choice tasks'; Briggs, Alonzo, Schwab, & Wilson, 2006). In our case, we want to offer the three levels of the 'model of model competence' in one task and none of them are wrong. Still, previous work showed that students hold additional perspectives in the aspect 'purpose of models' that are not included in the three levels (Grünkorn, 2014), therefore exhaustivity is not given and is not wanted to be achieved by artificially creating wrong

answers as this makes the interpretations of scores more difficult (Terzer, 2012). If exhaustivity is not to be achieved, Moosbrugger and Kelava (2012) recommend the following:

> "Bei der Konstruktion der Antwortalternativen von Multiple-Choice-Aufgaben im Persönlichkeits- und Einstellungsbereich ist die Beachtung der Exhaustivität der Alternativen besonders wesentlich. Sofern keine Exhaustivität vorliegt, kann mit einer entsprechenden Instruktion veranlasst werden, dass sich die Probanden für die am ehesten auf sie zutreffende Antwortalternative entscheiden, auch wenn keine der Optionen für den Probanden genau passt. Solche Antwortformate werden „Forced Choice" genannt und z. B. bei Interessenstests benutzt." (p. 49)[18]

We decided to use forced choice tasks where students have to choose one out of three given answer options, each representing one of the three levels from the 'model of model competence'. Thereby each student indicates "which of the [options] included in the item is most indicative of his or her behaviour" (McCloy et al., 2005, p. 225). This type of forced choice task, which we deem most easily interpretable, is a simplification of rank order tasks (cf. Cooper, 1983). In the latter, the answer options have to be brought into a preference rank order. Krell (2013) summarised considerations concerning the statistical interpretation of this forced choice task type and resulted in generating 'partially ipsative data' (Hicks, 1970) by merely scoring the most preferred level for each student in the rank order tasks. This was done due to the observation that the students' second ranked rating was "not content valid in some cases" (Krell, Reinisch, & Krüger, 2015).

[18] "When constructing answer options for multiple-choice tasks in the domain of personality and attitudes, the consideration of the exhaustiveness of the answer options is particularly important. If there is no exhaustiveness, an instruction can ask the test takers to choose the most appropriate answer option for them, even if none of the options fits exactly. Such response formats are called "forced choice" and are being used, for example, for interest tests."

4.2.3. Choice of contexts and construction of answer options

After picking the task format, it had to be decided whether the tasks are to be contextualised or decontextualised, resulting for this study in the choice of either using concrete models in the tasks or not.[19] In the field of research into students' meta-modelling knowledge, there are decontextualised (e.g., Grosslight et al., 1991; Treagust et al., 2002) as well as contextualised (e.g., Al-Balushi, 2011; Grünkorn, 2014; Krell, 2013) approaches. Many researchers question the validity of decontextualised approaches to assess students' meta-modelling knowledge (Grünkorn, Upmeier zu Belzen, & Krüger, 2014; Krell et al., 2012; Sins et al., 2009). For example, it is being argued that students have context-bound epistemological ideas which they may not be able to generalise. Sins et al. (2009) state:

> "However, the approach taken by Grosslight et al. (1991) focuses on an epistemology of science at a rather abstract level. The validity of this approach has been questioned, because students' epistemological ideas are likely to be context-bound (e.g., Elby & Hammer, 2001; Hofer & Pintrich, 1997; Hogan, 2000; Paulsen & Wells, 1998). As a consequence, in specific contexts, students may entertain advanced epistemologies without being able to articulate the underlying viewpoints in general." (p. 1208)

Other researchers say that a general view on models may not be the aim after all. Nehm and Ha (2011) emphasise that contextuality is an important issue in biology more than in other science domains because "the core units of biology – individuals and species – are not conceptualised by scientists as structurally or compositionally the same in different places or in different

[19] The term 'context' is being used broadly in the field of science education, covering (1) contexts as task features (e.g., employing different models; Krell et al., 2012; Al-Balushi, 2011), (2) contexts as discipline-specifications (e.g., tasks in biology, chemistry, physics; Gobert et al., 2011; Krell et al., 2015) and (3) contexts as learning situations (e.g., support in learning; Berland and Crucet, 2016; van Oers, 1998). I refer to the term 'context' meaning different models as features in the task stems.

environments" (p. 239). It follows from this that it should be required of students to be able to express their epistemological ideas within specific contexts and be aware of different purposes regarding different models (e.g. school models vs. scientific models; Gilbert et al., 2000; Treagust et al., 2002; Trier et al., 2014). More and more authors request researchers to develop contextualised instruments (Al-Balushi, 2011; Krell et al., 2012; Lee et al., 2015). Lee et al. (2015) emphasise clearly:

> "As mentioned previously, one of the major implications of the aforementioned results indicated the importance of including context and examples of a variety of models in future questionnaires. […] While models vary and the roles of models in science are complex and heterogeneous across different disciplines, topics, or phenomena, we believe it is important to develop more context-based instruments to provide more careful and contextualized interpretations of the results." (p. 21)

Following this request, contextualised tasks were developed. Decisions for the choice of contexts and an exemplary forced choice task for the aspect 'nature of models' have been described in Article 1:

Gogolin, S., & Krüger, D. (2015). Nature of models – Entwicklung von Diagnoseaufgaben. In M. Hammann, J. Mayer, & N. Wellnitz (Eds.), ***Lehr- und Lernforschung in der Biologiedidaktik*** **(6th ed., pp. 27–41). Innsbruck: Studienverlag.**

For the selection of contexts, we used the experiences and accomplishments of previous test development projects in the same field (Grünkorn, 2014; Krell, 2013; Terzer, 2012). In order to determine the representational mode of the models to be used in the tasks, we adapted the distinction between concrete and abstract models (Coll, 2006; Suckling et al., 1978), which was also determined empirically by Krell et al. (2014b) in a student-based typology of biological models.

Considering content-related aspects, we tried to include a variety of contexts from different areas such as physiology, ecology and evolution to improve content-representation (Hartig & Frey, 2012; Messick, 1995). We did not include models of genetics as experts reported that in their previous research, students had problems understanding the content of these models and thus focussed rather on the content than on the model (Fleige et al., 2012). Instead of trying to equally cover all content areas of school books, we tried to find contexts which allow the construction of answer options on all three levels of the 'model of model competence' that are accessible and plausible for students (Grünkorn, 2014). For this to be the case, the contexts needed to be realistic and allow recognising the hypothetical nature of the original. Following Giere (2001), we found contexts useful which refer to originals that cannot be accessed directly due to restrictions of time (e.g., original is too remote or too complex) or size (e.g., original is too small or too large to be directly observed with the human eye). Pictures of all contexts used during test development, including information about their stance on the concrete-abstract continuum and about the licences, are assembled in the Appendix (Attachment 2). For each of the contexts, we used the theoretical descriptions of each level from the 'model of model competence' and corresponding student answers from the previous study by Grünkorn (2014) to construct, on every level, two or three answer options (cf. Table 5).

Table 5. Construction of answer options for the three levels in the aspect 'nature of models' from the 'model of model competence' (Grünkorn et al., 2014; Krell et al., 2016).

	Level I	**Level II**	**Level III**
Theoretical description	Model is a replication of the original	Model is an idealised representation of the original	Model is a theoretical reconstruction of the original

Student answer (Grünkorn et al., 2014)	*"They had sharp teeth and a long, well-built tail. They had two big legs to walk and two short arms."*	*"The model only shows the essential parts […]. The main traits, structures and colours are shown here."*	*"Because one cannot know what a bio membrane really looks like. The thin lines look similar to the model but that's it. The scientists only assume."*
Constructed statement	The model of the T. rex shows that the real T. rex had short arms, big teeth and a long tail.	The model of the water cycle shows the essential stages of the water cycle because the scientists left out some minor steps.	The model of the bio membrane may look like a real bio membrane, but one can only assume what it really looks like.

In the resulting forced choice tasks, three answer options were to be assembled (Table 6). A task stem shows a biological model with a short description (in the aspect 'nature of models', the task stem additionally shows the original which the models refers to).

Table 6. Scheme for the construction of forced choice tasks for the aspects 'nature of models' and 'purpose of models'.

		'Nature of models'	**'Purpose of models'**
Task stem	**Introductory phrase**	On the left, there is [X] and on the right, there is a model [X] which was made by scientists.	On the picture, there is a model [X] which was made by scientists.
	Picture	Picture of [X] / Model [X]	Model [X]
Standardised stimulus		**State to what extent this model** [X] **corresponds to** [X] **that occurs in nature.**	**Models are being made for a certain purpose. State, which purpose this model** [X] **may serve.**
Answer options (AO)	**Introduction**	**The model** [X] **…**	**The model** [X] **permits to ...**
	Body	… AO Level I … AO Level II … AO level III (in randomised order)	… AO Level I … AO Level II … AO level III (in randomised order)

We decided to include as little content-related information as possible in the task stem in order to avoid construct-irrelevant variance (Messick, 1995).

Nevertheless, we did include the phrase "which was made by scientists" as construct-relevant information because Trier et al. (2014) found the distinction between the school and the science context to have an influence on students' answers. The task stem is followed by a standardised stimulus and three answer options, each representing one of the three levels of the theoretical framework. We tried to avoid inequities by keeping the text length as even as possible across answer options, by using the same level of language (e.g., no foreign words) in each of the answer options, by not using double negatives or any other terms that may be especially compelling to the students in one answer option and not in another (Döring & Bortz, 2016; Moosbrugger & Kelava, 2012; Raymond, Lane, & Haladyna, 2015). These aspects were later to be reviewed in an expert judgement.

4.2.4. Expert judgement

In this research project, the created answer options for both aspects ('nature of models' and 'purpose of models') were investigated into by means of expert judgement studies. This chapter summarises the studies I and V from Table 3. The results of these studies have not been published before. In the expert judgement studies, the initially developed tasks for the aspects 'nature of models' ($N1_9$) and 'purpose of models' ($P1_6$) were given to a panel of experts (researchers in the field of biology education) in order to gain evidence for validity regarding construct representation, the latter being a part of 'test content' (AERA et al., 2014; Messick, 1995).[20]

[20] The expert judgements served as the very first check of the initial six contexts for the aspect 'nature of models' (TR, NT, BM, VS, EV, AR) with three answer options per level (= 9 answer options per context) and the initial eight contexts for the aspect 'purpose of models' (TR, JW, BM, BR, EV, LZ, AS, BG) with two answer options per level (= 6 answer options per context). The initially created answer options ($N1_9$ and $P1_6$) are listed in chapter 10.5 in the Appendix.

RQ Studies I and V — To what extend is the experts' rating of the levels of the answer options consistent with the theoretically intended level of the answer options?

In order to adequately represent the content domain, an experts' rating of the level of the answer options should match with the theoretically intended level of the answer options.

In study I, eleven experts gave feedback on the tasks in the aspect 'nature of models' and in study V, nine experts gave feedback for the tasks in the aspect 'purpose of models' respectively. In both studies, the online platform 'SoSci Survey' was used to question the experts. In the questionnaire, the experts had to match every answer option of each of the contexts with one of the three levels of the 'model of model competence' (Screenshot of the expert judgement questionnaire see Attachment 3), thereby indicating whether the answer options adequately represent the construct to be measured (Messick, 1995). The 'model of model competence' was made available to the experts, as the judgements are known to improve when firmly based on theory (Raymond et al., 2015). For every context, there were some additional questions taken from Kauertz, Löffler, and Fischer's (2010) 'key questions for the critical reflection of tasks' which gave the experts the chance to formulate a more general feedback for the task or to give further justification for their decisions (Attachment 4).

During data analysis, the fit between the theoretically intended level of an answer option and the level attributed to this answer option by the experts was determined by using cross tables and by calculating Cohen's Kappa (κ) with the levels being treated as nominal data with no intrinsic order. Rather than for calculation of classical interrater-reliability, Kappa was in this case used to "describe the degree of agreement between an assigners' judgement about objects and the known criterion status of these objects" (Brennan &

Prediger, 1981, p. 689). Figure 2 and Figure 3 show the fit between experts' interpretation and theoretical intention for all answer options.

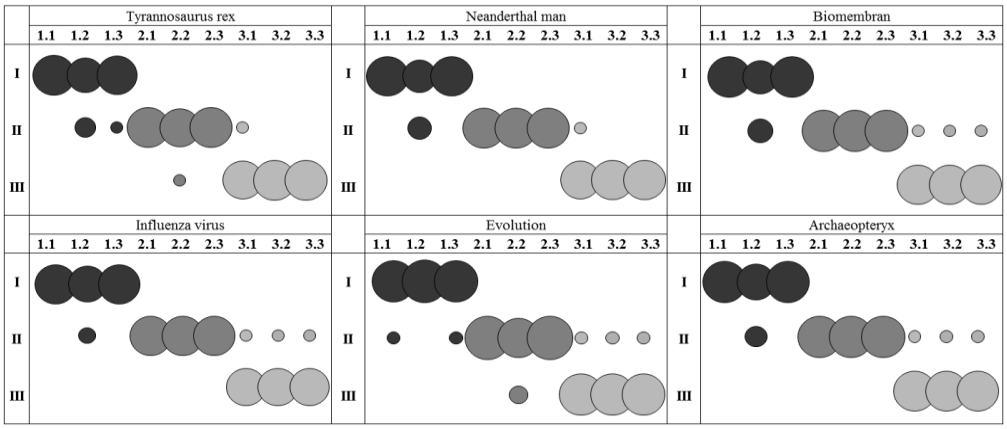


Figure 2. Bubble chart of the expert judgement (N = 11) for the aspect 'nature of models'. The columns represent the theoretically intended level of the answer options and the rows are the judgements of the experts as to which level the answer option belongs. Dark grey = Level I; middle grey = Level II; light grey = Level III.

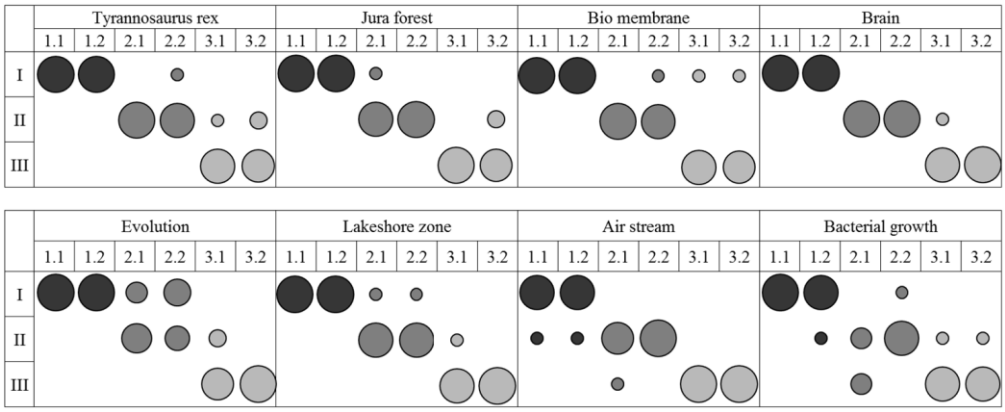


Figure 3. Bubble chart of the expert judgement (N = 9) for the aspect 'purpose of models'. The columns represent the theoretically intended level of the answer options and the rows are the judgements of the experts as to which level the answer option belongs. Dark grey = Level I; middle grey = Level II; light grey = Level III.

The cross tables as well as the qualitative feedback by the experts were discussed with some of the experts in colloquium sessions at the Freie

Universität Berlin. For example, in the aspect ‘nature of models’, answer option 1.2 (e.g., “The model looks like a real Neanderthal man because I imagine the shape of the eye brows and the nose to have been similar.”) which refers to the category ‘Model represents a (non-) subjective conception of the original’ by (Grünkorn et al., 2014) has been criticised by some experts as not clearly distinguishable from the perspective of seeing a model as a theoretical reconstruction:

> „Ich frage mich aber, warum subjektive Vorstellungen des Probanden („ich“) weniger bedeuten als subjektive Vorstellungen von Wissenschaftlern („… ähnelt dem damals lebenden Neandertaler möglicherweise, Wissenschaftler können aber nur vermuten und nicht wissen, wie er ausgesehen hat.“).“[21]

For this reason, answer option 1.2 was consensually rephrased into “The model of the Neanderthal man shows that the Neanderthal man had strong eye brows, a broad nose and a receding chin.”. Further consequences for the development of answer options in both aspects are summarised in Table 7.[22]

Table 7. Consequences of the expert rating for the development of answer options.

‘nature of models’	‘purpose of models’
– Adaptation of foreign or signal words within all answer options (if necessary)	– Replacement of the verb “verdeutlichen” by “erklären”
– Rewording of answer option 1.2	– Replacement of the verb “untersuchen” by “Vermutungen ableiten”
– Reduction to two answer options per level for the context EV	– Replacement of the verb “erkunden” by “vorhersagen”

[21] “I wonder why subjective perceptions of the test taker ("I") are worth less than subjective perceptions of scientists (“... may resemble the Neanderthal man, but scientists can only persume and not know what he looked like.”).”

[22] The optimized answer options (after the expert judgement; $N2_{6\text{-}9}$ and $P2_6$) are to be found in chapter 10.5 in the Appendix.

– Design of a new context (WC) similar to the context of EV	– Refinement of the verb “erforschen” by “weiter erforschen”
– Design of two new contexts (AS, WM) that are more abstract so to cover the spectrum from concrete to abstract more broadly	– Deletion of the term “evolution” for the level II answer option of the context EV
– Dismissal of the AR context	

With reference to the research question, the Kappa values ($0.72 < \kappa < 0.94$) show that there is a very good fit between experts’ rating of the levels of the answer options and the theoretically intended levels (Brennan & Prediger, 1981). This indicates an adequate representation of the content domain (Messick, 1995). The expert judgement also helped to bring to light some representational weaknesses in the initial answer option formulations which may have jeopardised the proposed score interpretations later (Messick, 1995). The own experiences in task development by some of the experts helped to anticipate problems students would be likely to have with certain words or phrasings. Nevertheless, Messick (1984, 1995; Messick) repeatedly mentions that any form of expert judgement is fallible and should be supported by other kinds of evidence. Leighton (2004) emphasises that “adult [expert] inferences, however, can sometimes fail to catch problematic item features.” (p. 10). Although some problematic formulations may have been avoided by this procedure, gathering evidence for validity based on response processes by the actual target test takers is indispensable in order to verify the answer options’ adequate representation of the construct empirically.

4.3. Evidence Based on Response Processes

At the end of the first step of task development, including the expert judgements for both aspects respectively, we had optimised answer options which now had to be put to the test with students. Although the integration of students in the process of test development has been demanded repeatedly (Adams & Wieman, 2011; AERA et al., 2014; Leighton, 2004; Leighton & Gierl, 2007a), it is often not being realised due to the high expense of resources (Leighton, 2004). Leighton (2004) refers to Baxter and Glaser (1998) to say that we need to "take into account how students interpret and respond to test item features, serving as a check of whether the expected knowledge and skills measured by test items are indeed exhibited by students." (Leighton, 2004, p. 8). Floyd et al. (2005) name 'Evaluation of instrument instructions and response formats', 'Interviews with informants about thought processes' and 'Think aloud protocols with informants' as possible sources to gather evidence based on response processes. In this chapter, I will summarise the findings of three studies which included students in the process of test development. In study IV, we obtained concurrent think aloud protocols and descriptions of thinking for the tasks in the aspect 'nature of models'. In study VI, we performed a mixed methods study where the answer options in the aspect 'purpose of models' were implemented in Likert-scales and subsequent student interviews were performed.

4.3.1. Concurrent think aloud protocols and descriptions of thinking

Think aloud protocols can provide information concerning students' perspectives on tasks, which cognitive processes are being triggered and which strategies are being used to answer the tasks (Ericsson & Simon, 1998; Gogolin & Krüger, 2016a; Knoblich & Öllinger, 2006; Leighton, 2004). In a think aloud protocol, a student is told to think aloud while working on a particular problem solving task. In response to Nisbett and Wilson (1977) who criticised that think aloud studies produce inaccurate data, Ericsson and Simon (1980) distinguished between 'concurrent' and 'retrospective' thinking aloud. Concurrent thinking aloud provides direct verbalisations of cognitive processes, thus being believed to be consistent and complete. Retrospective thinking aloud (also called 'description of thinking') results from the interviewees being asked to recall what they were thinking during the work on a task after the task had been completed, thus being more susceptible to thought-altering influences (Ericsson & Simon, 1998; Fonteyn, Kuipers, & Grobe, 1993; Konrad, 2010; Sandmann, 2014). Ericsson and Simon (1993) further distinguish between 'talk aloud' (vocalisations of thoughts that are already encoded in the verbal form), 'think aloud' (verbalisation of a sequence of thought that are held in memory in some other form, e.g. visually) and 'other verbalisations' (retrospective reports on thoughts not held in memory). We conducted both concurrent think aloud protocols and retrospective descriptions of thinking.[23] Some of the results of the concurrent think aloud protocol analysis have been published in Article 2:

Gogolin, S., & Krüger, D. (2016a). Diagnosing Students' Understanding of the Nature of Models. ***Research in Science Education,*** **1–23. doi: 10.100 7/s11165-016-9551-9.**

[23] To be exact, it was a 'talk aloud' study because the students' referred to thoughts that were encoded in the verbal form. Still, the term 'think aloud' is nowadays used as an umbrella term.

These results are to be complemented in this dissertation by the so far unpublished data from the retrospective description of thinking protocols and the joint analysis before the background of validity evidence. In order to gain evidence for validity regarding students' response processes (AERA et al., 2014; Leighton, 2004; Messick, 1995), the following research question was formulated:

RQ $_{\text{Study IV}}$ To what extend is the students' interpretation of the answer options for the aspect 'nature of models' consistent with the theoretically intended level of the answer options?

Analogous to the expert judgement studies (I and V), the students' interpretations of the levels of the answer options in the forced choice tasks should match with the theoretically intended level of the answer options in order to support the validity argument. In the study, 30 students from grades nine to twelve of Berlin secondary schools concurrently thought aloud while answering eight forced choice tasks ($N3_6$).[24] Immediately after, we followed up with retrospective descriptions of thinking. Each student was given the six answer options of one or two contexts again and was asked to describe in own words how he or she understood every single answer option. The concurrent think aloud protocols ($N_{\text{answers}} = 254$) and the retrospective description of thinking protocols ($N_{\text{answers}} = 180$) were analysed by two independent raters who, for each answer option, rated the students' responses binary in the categories 'understood as intended' (1) or 'not understood as intended' (0).[25] Responses that could not clearly be interpreted by the raters, were equally rated 'not understood as intended' (0). Cohen's Kappa (κ) was computed as a measure of agreement between raters ($\kappa_{\text{Interrater}} = .81$; very good; Wirtz & Caspar, 2002). In order to select

[24] The instructions for the interviewer can be found in Attachment 5.
[25] Coding scheme is in the Appendix, Attachment 6.

answer options for the final forced choice tasks, we chose the answer option per level and per context that was most frequently understood as theoretically intended by the students. This selection was based solely on the responses by students from grades eleven and twelve as a grade-specific analysis had revealed that the answer options were not understood adequately by students from lower grades (see chapter 4.5.1. 'Fairness in testing for subgroups'). Table 8 shows the findings of the combined analysis of the concurrent think aloud protocols and the description of thinking protocols and indicates which of the answer options were selected.

Table 8. Students' responses [%] that were rated as understood per context, per level and per answer option in the aspect 'nature of models' based on eleven and twelve grade students' responses in concurrent think aloud protocols ($N_{answers}$ = 117) and in retrospective descriptions of thinking ($N_{answers}$ = 92).

		Level I		Level II		Level III	
		Ia	Ib	IIa	IIb	IIIa	IIIb
T. rex	[TR]	57 (7)	100 (4)	67 (6)	75 (4)	25 (4)	100 (3)
Neanderthal man	[NT]	67 (3)	100 (2)	80 (5)	71 (7)	89 (9)	100 (4)
Bio membrane	[BM]	100 (3)	100 (2)	80 (5)	75 (4)	60 (5)	60 (5)
Influenza virus	[VS]	100 (2)	100 (1)	100 (6)	75 (4)	50 (6)	50 (4)
Evolution	[EV]	75 (5)	100 (4)	100 (5)	100 (5)	50 (6)	100 (4)
Water cycle	[WC]	67 (6)	88 (8)	80 (5)	100 (5)	50 (2)	67 (3)
Air stream	[AS]	100 (2)	100 (3)	100 (6)	75 (8)	0 (3)	0 (4)
Water melon	[WM]	100 (2)	100 (1)	100 (8)	83 (6)	0 (3)	0 (1)

Note: Grey: Answer options that were finally selected. Little number in brackets indicates the number of analysed answers.

Table 8 shows that the students' correct interpretation of the answer options decreases with higher level for some contexts. Also, the contexts 'Air stream' and 'Water melon' are not being understood sufficiently as theoretically intended. By producing such patterns, the protocol analysis

helped to select answer options for the final forced choice tasks in the aspect ‘nature of models’ based on students’ adequate responses. Still, the transcription and interpretation of the concurrent think aloud protocols was time consuming and, even in combination with the description of thinking protocols, only a very limited amount of student responses could be obtained per answer option (1-9 responses). Possible reasons for the level- and context-specific findings as well as the methodical constrains of this approach are being discussed with reference to their impact on the validity argument in chapter 5.2.

4.3.2. Mixed methods study with rating scales and student interviews

We used the experiences made in the task development process for the aspect ‘nature of models’ to improve the efficiency and informative power of the procedure for the aspect ‘purpose of models’. The following chapter will present the two-step procedure of the selection of answer options for the tasks in the aspect ‘purpose of models’ based on students’ responses. It is a summary of Article 3:

Gogolin, S., & Krüger, D. (2016). Konstruktion von Diagnoseaufgaben zum Zweck von Modellen. *Biologie Lehren und Lernen - Zeitschrift für Didaktik der Biologie, 1*(20), 44-62.

We used a mixed methods procedure for test development which is both based on students’ (N = 275; grades 9-12) decisions on rating scales and on retrospective student interviews (convergent mixed method design; Creswell & Plano Clark, 2011). In this study, we used the answer options for the aspect ‘purpose of models’ which had already been examined by the experts

($P2_6$). In order to further examine and to select suitable answer options for the forced choice tasks, two research questions were investigated empirically:

RQ1 $_{\text{Study VI}}$ To what extent is the students' interpretation of the answer options for the aspect 'purpose of models' consistent with the theoretically intended level of the answer options?

RQ2 $_{\text{Study VI}}$ To what extent differ the answer options for the aspect 'purpose of models' within each context in their relevance for students?

The first research question is analogous to the one in the think aloud study for the aspect 'nature of models'. Nevertheless, instead of performing concurrent think aloud protocols, we now retrospectively asked students in fully structured interviews why they rated each answer option the way that they did. In order to provide evidence for validity, students' interpretations of the answer options would need to be consistent with the theoretically intended level (AERA et al., 2014; Leighton, 2004; Messick, 1995). To gather empirical student based evidence for adequate construct representation, we investigated the second research question concerning the relevance that students attribute to the content-related differences across answer options within one context. We expected content-related differences to have an impact on how relevant the answer options were perceived by students (cf. Cohors-Fresenborg, Sjuts, & Sommer, 2004; Kauertz, 2008; Krell, Upmeier zu Belzen, & Krüger, 2014a). For the assembly of the forced choice tasks, we were interested to find those answer options which were understood correctly by the students and which were most relevant to students in relation to the given context.

In a first step of the selection procedure, we analysed the retrospective fully structured interviews with the students (n = 160 students; 1920 statements total; 40 statements per answer option) by means of an evaluative content analysis (Kuckartz, 2016) deciding whether each answer option was understood as intended or not.[26] It was possible to find at least one answer option per level per context, which we claim will be understood as intended by the students, except for the contexts 'Air stream' and 'Bacterial growth'. Were there two answer options per level remaining for a context, we additionally compared the students' ratings concerning differences in their relevance for the students. For all contexts and levels, we chose the answer options that were both understood sufficiently as intended by the students and that had the highest value of relevance on the rating scale.

Using the procedure, we included students' responses to select answer options for forced choice tasks for six of the initial eight contexts. Although all answer options had before been approved of by experts, the students' responses revealed imprecisions and content-related imbalances across answer options. Due to the combination of quantitative (rating scale) and qualitative measures (student interviews) within a convergent mixed method design (Creswell & Plano Clark, 2011), it was possible to include a large number of student responses without sacrificing too much resources for data collection.

[26] Coding scheme is enclosed in the Appendix, Attachment 7.

4.4. Evidence Based on Relations to Other Variables

Evidence based on relations to other variables can be gathered by comparing the test under investigation with external variables, e.g., scores from another test which is either intended to measure the same or a different construct (AERA et al. 2014). The convergence of independent methods measuring the same trait indicates evidence for validity (convergent validity; Campbell & Fiske, 1959), whereas the divergence of independent methods measuring different traits likewise indicates evidence for validity (discriminant validity; Campbell & Fiske, 1959). Also, it ought to be demonstrated whether a test can discriminate between groups known to differ on the variable of interest.

A practical approach to gathering evidence based on relations to other variables is the multitrait-multimethod (MTMM) paradigm (Campbell & Fiske, 1959). The MTMM paradigm assumes to measure each of several concepts (Campbell & Fiske, 1959) by each of several methods (e.g., closed-ended tasks, open-ended tasks, observation). Moosbrugger and Kelava (2012) summarise Campbell and Fiske's (1959) MTMM to traditionally be a correlation matrix between measures with estimates of reliability in the diagonal of each measure (monotrait-monomethod). Correlations among measures of the same trait which were obtained using different methods constitute the validity diagonal (monotrait-multimethod). Correlations among measures that share the same method of assessment constitute the multitrait-monomethod triangles and give an indication of the method bias. Multitrait-multimethod triangles differ in both trait and method. Since the work by Campbell and Fiske (1959), evidence for validity based on relations to other variables askes for the investigation of both convergent (e.g. by monotrait-multimethod analysis) and discriminant (e.g. by multitrait-multimethod analysis) validity. Consequently, an instrument

measuring meta-modelling knowledge should correlate highly with another instrument that also measures meta-modelling knowledge and at the same time it should correlate only little with an instrument measuring, for example, numeracy.

In the course of this research project, we pursued two approaches based on relations to other variables. In study III, we conducted a monotrait-multimethod study with the $N3_6$ version of our forced choice tasks, with open-ended diagnostic tasks and with diagnostic interviews. We also performed an intervention study (study VIII) that was designed to check whether the instrument is sensible enough to track changes in students' meta-modelling knowledge when being used before and after a promotion of model competence. Evidence for validity would be obtained in this study if the theoretically expected pattern (improvement of meta-modelling knowledge) could actually be identified in the data (better scores). Both studies are to be summarised in the following subchapters.

4.4.1. Monotrait-multimethod approach

Article 2 reports design and findings of study III and discusses the evidence based on relations to other variables with regard to validity.

Gogolin, S., & Krüger, D. (2016a). Diagnosing Students' Understanding of the Nature of Models. *Research in Science Education,* 1–23. doi: 10.100 7/s11165-016-9551-9.

In the study, three independent methods were used to measure the students' (N = 448) meta-modelling knowledge (monotrait-multimethod) in the aspect 'nature of models' and thereby produced evidence that helps to judge upon convergent validity.

RQ $_{\text{Study III}}$ To what extent do students have a consistent understanding of the nature of models when being assessed with forced choice tasks, open-ended diagnostic tasks and diagnostic interviews?

As mentioned before, we expect the students' scores for all diagnostic methods to converge in order to indicate convergent validity (Campbell & Fiske, 1959).

Students' meta-modelling knowledge in the aspect 'nature of models' was diagnosed using twelve forced choice tasks ($N2_{6\text{-}9}$), two open-ended diagnostic tasks and an interview with two diagnostic questions ($N = 448$; $n_{\text{interview}}=194$; grades 7-12).[27] The students' responses to the open-ended tasks and in the interviews were coded by the means of an evaluative content analysis (Kuckartz, 2016). One of three levels was assigned to each response based on the descriptions of the aspect 'nature of models' of the 'model of model competence' (Krell et al., 2016). For the comprehensive analysis of the data, we estimated a partial credit model (Masters, 1982) modelling the methods as three latent dimensions. Students' meta-modelling knowledge was inferred for each of the three dimensions using Warm's (1989) Weighted Likelihood Estimator (WLE). On the basis of the WLEs, grade-specific dependent *t* tests were performed to examine whether the students answered differently depending on the diagnostic method.

The three-dimensional partial credit model fitted the data of students' meta-modelling knowledge appropriately. The item separation reliability of 0.925 was excellent (Bond & Fox, 2001), whereas the person reliability (EAP/PV)

[27] The instructions for the survey procedure can be found in Attachment 8. Open-ended tasks / interview questions: "Beschreibe, inwieweit ein von Biologen entwickeltes Modell seinem biologischen Original entspricht." [State to what extent a biological model which was made by scientists corresponds to its original.] and "Beschreibe, inwieweit sich ein von Biologen entwickeltes Modell vom biologischen Original unterscheidet." [State to what extent a biological model which was made by scientists differs from its original.].

was similarly low for all three assessment methods (twelve forced choice tasks: 0.508, two open-ended tasks: 0.550, two interview questions: 0.538). The grade-specific dependent *t* tests showed that in grades 7/8 and 9/10, the students were accredited abilities for the open-ended tasks and the interviews that were significantly lower than for the forced choice tasks. In grades 11/12, the students did not answer significantly different across the open-ended tasks, the interviews and the forced choice tasks. A more detailed account of the findings with regard to the grade-specific differences of students' meta-modelling knowledge and possible method-specific influences can be found in Article 2 (chapter 11) and will be referred to again in chapter 4.5.1. ('Fairness in testing for subgroups').

In order to provide convergent validity evidence, we expected the diagnostic methods to converge in the measurement of students' meta-modelling knowledge (Campbell & Fiske, 1959; AERA et al., 2014). A conversion of scores was found for the subgroup of students from grades eleven and twelve.

4.4.2. Intervention

> "There are several practical requirements for such an instrument if it is to be used on a widespread basis so it can impact instruction. (1) It must measure value added by the instruction, and hence it must be possible to administer it on both pre- and post-instruction basis." (Adams & Wieman, 2011, p. 1292).

The detection of even small gains in understanding is necessary when wishing to sensibly use diagnostic instruments in class (Hartig et al., 2008). For the students' scores to be interpreted validly with regard to this purpose, we need to gather evidence that the instrument is sensitive enough to track changes in students' meta-modelling knowledge after an intervention.

RQ $_{\text{Study VIII}}$ Does the diagnostic instrument track changes in students' meta-modelling knowledge in the aspects 'nature of models' and 'purpose of models' over the course of an intervention?

Evidence based on relations to other variables relies on the comparison of the test under investigation with external variables to demonstrate whether a test can discriminate between groups theoretically expected to differ on the variable of interest. Consequently, we predicted that students, after the intervention, should have an improved meta-modelling knowledge in the aspects 'nature of models' and 'purpose of models' (AERA et al., 2014).
We conducted an intervention study (study VIII) with 65 students (four biology A-level courses) in a special exhibition ('modellSCHAU'; Grotz, 2015) in the Botanical Garden Berlin. All of the students completed the twelve tasks (six for each aspect) before and after the 60 min long intervention. The stages of the intervention as well as the results of the study have been described in Article 5:

Gogolin, S. & Krüger, D. (in press). Modellverstehen im Biologieunterricht diagnostizieren und fördern. *Der mathematische und naturwissenschaftliche Unterricht.*

The results show that the students' meta-modelling knowledge improved in both aspects in the course of the intervention. The Wilcoxon test attests the differences in both aspects to be statistically significant ($p_{Nature} < .000$; $p_{Purpose} < .000$; $r_{Nature} = .563$; $r_{Purpose} = .490$). The findings of the intervention study support the theoretical expectation of an improved meta-modelling knowledge and can therefore be interpreted as evidence of validity based on relations to other variables (AERA et al., 2014). Methodical constrains of this approach will be discussed in chapter 5.3.

4.5. Evidence Based on Internal Structure

In the modern concept of validity, evidence based on internal structure contains a number of aspects that used to be treated separately, e.g., the issues of reliability and fairness in testing (chapter 2.2). The inclusion of these aspects into the validity argument goes back to Messick (1995) who names construct-irrelevant variance (e.g., the negative affection of some individuals because of characteristics irrelevant to the intended construct) as one of two major threats to validity. In order to minimise construct-irrelevant variance, the Standards for Educational and Psychological Testing ask, among other things, to consider whether valid inferences from the results can be drawn for all students, or only for some subgroups but not for others (AERA et al. 2014). We therefore paid respect to grade-specific differences in most of our investigations. Some of our attempts will be delineated in the subchapter *Fairness in testing for subgroups.*

Another aspect of the internal structure is the variability allowed in the testing procedure. With respect to the aspect of reliability, the Standards for Educational and Psychological Testing highlight that, if the interpretation of the scores assumes that the construct being assessed does vary over occasions, the variability over occasions needs to be respected in the interpretation of scores, otherwise being a potential source of measurement error (AERA et al. 2014). We investigated into context-specific differences in students' answers to our instrument and analysed the dimensionality of our measurement. The findings are being presented in the subchapter *Reliability*.

The original notion of objectivity is still present in the Standards for Educational and Psychological Testing as the aspect 'Fairness in Treatment during the Testing Process' (AERA et al. 2014). This aspect contains administration conditions and scoring procedures. It is also referred to by

Nichols (1994) in the steps ‘Test administration’ and ‘Response scoring’. In the subchapter *Diagnostic procedure*, I will thus elaborate on the final online version of the instrument and the score interpretation algorithm.

4.5.1. Fairness in testing for subgroups

In the theoretical frame, I argued that standards documents have called for an increased role for models in science classrooms. Model competence being part of the section “Erkenntnisgewinnung” (KMK, 2005) and clearly demarcated in the Berlin framework curriculum (SBJS, 2015) initially lead us to develop an instrument that can diagnose students from grades seven to twelve in Berlin secondary schools (‘Gymnasium’). Ultimately, we arrived at recommending the tasks in the aspect ‘nature of models’ to be used in grades eleven and twelve and the tasks for the aspect ‘purpose of models’ in grades ten to twelve. These recommendations are based on a number of studies that respected grade-specific differences of the students. In the aspect ‘nature of models’, the combined interpretation of study III and study IV lead to the exclusion of grades seven to ten. The embedded analysis of both of these studies has been reported in Article 2 (Gogolin & Krüger, 2016a). As the article mainly refers to gathering evidence for validity based on relations to other variables, it has been summarised in chapter 4.4 under this perspective. Still, the analysis was being performed and reported for separate subgroups of students which eventually lead to the exclusion of some of the subgroups and thus the article is also being referred to here. In the aspect ‘purpose of models’, the mixed methods study with rating scales and student interviews (study VI) that was reported in chapter 4.3.2., did not reveal differences in understanding across grades so the tasks were initially recommended for grades nine to twelve. Nevertheless, in study VI, the

students did not solve the final forced choice tasks but merely rated the single answer options on a scale.

In order to check again whether all subgroups of students understand the final tasks as intended, which is a precondition for a sound score interpretation, we decided to include open-ended justification tasks as a measure of control in our main study with the final diagnostic tasks for both aspects (study VII).[28] We investigated the following research question with regard to subgroups of students from different grades:

RQ $_{\text{Study VII}}$ To what extent is students from different grades' interpretation of the forced choice tasks consistent with the theoretically intended content of the tasks?

In each of the two open-ended tasks, the students were asked to give reason for their decision in one of the forced choice tasks:

> "You just answered six questions concerning different biological models. For the model [X], please give reason for why you chose your answer options. Please also explain why you didn't choose the other answer options."

In contrast to think aloud protocols or interviews, the use of these open-ended justification tasks allowed for a greater sample to be included (Adams & Wieman, 2011). Altogether, we included 720 written responses (20 justifications per forced choice task / per grade). The responses were used to verify if all subgroups of students chose their answer options for legitimate and logically sound reasons (cf. Treagust et al., 2004). The responses were transcribed verbatim and analysed with an evaluative content analysis (Kuckartz, 2016). In this procedure, the responses were coded and scaled

[28] A more detailed description of the study design is to be found in Article 4 (chapter 11) and in chapter 4.5.2., where the findings are summarised with regard to the issue of reliability.

based on four pre-existing categories (Attachment 10). Either the student did or did not understand the chosen answer option according to theory or the student used a strategy to pick an answer option, for example by guessing or by excluding all other answer options. The response was ‘not rateable’ if there was no indication in the explanation as to why the student decided for a certain answer option. The categories were assigned to the students’ responses based on a coding scheme which contained descriptions of the aspects ‘nature of models’ and ‘purpose of models’, coding rules and examples of student responses for each of the codes. In order to judge the agreement between raters, another independent rater double coded 50 % of the material and Cohen’s Kappa (κ) was computed (Wirtz & Caspar, 2002). The interrater-reliability was very good for all tasks ($.78 \leq \kappa_{interrater} \leq .89$; Wirtz & Caspar, 2002).

Article 4 quantifies the responses depending on the chosen level of understanding in the forced choice tasks for the aspects ‘nature of models’ and ‘purpose of models’ depending on the grade of the students across all contexts. All in all, most students justified their choice in the forced choice tasks in a way that made us believe they understood the answer option and opted for it due to logically sound reasons. For example, for the aspect ‘nature of models’ in grade eleven, 105 out of 120 students were categorised as ‘understood’ while only seven were categorised as either ‘not understood’ or ‘strategy’. In grade twelve on the other side, ten students did not understand the answer option correctly. It has to be pointed out that a further in depth analysis of the student explanations revealed that seven of the ten responses categorised as ‘not understood’ were caused by students who interpreted the answer option on a higher level. The opposite is true for grade nine students’ explanations for the aspect ‘purpose of models’. A total of 18 students were categorised as either ‘not understood’ or ‘strategy’. 13 of these students had chosen an answer option on level III by strategy or by misinterpretation.

4.5.2. Reliability

Study VII, as it was the main study with the final diagnostic forced choice tasks from both aspects, also served for statistical analyses concerning reliability and dimensionality. As test developers, we need to have some indication of the reliability of the scoring procedure in order to evaluate the measured behaviour (AERA et al., 2014). As mentioned earlier, classical test theory recognises primarily 'alternate-form coefficients', 'test-retest coefficients' and 'internal-consistency coefficients' (Field, 2013). Cronbach's alpha, which is an internal-consistency coefficient, is the most common measure of scale reliability with a generally acceptable value of above .7 (Field, 2013). Before performing reliability analyses, it is important to define the dimensionality of the construct of interest. Cronbach's alpha, for example, measures the extent to which the scale measures one underlying factor or construct (Field, 2013). Changes in scores from one occasion to another (which would decrease reliability) are not regarded as error if they result from changes in the construct being measured (AERA et al., 2014). The dimensionality of our construct was being investigated on two levels. First, we checked whether the tasks for the two aspects ('nature of models' and 'purpose of models') display two distinct dimensions as theoretically expected (Krell & Krüger, 2011). Second, we investigated into differences across contexts within the single aspects, because there is evidence that students' meta-modelling knowledge may be contextualised. This would mean that there may be sub dimensions within each of the aspects, thus requiring multiple tasks for each sub dimension again in order to establish reliability satisfyingly. Before summarizing Article 4, where reliability measures and the question of contextuality are being discussed, I would like to present the unpublished

analysis of dimensionality for the two aspects of meta-modelling knowledge that were discriminated here.

As a basis for the dimensionality analysis, we had the answers of eleven and twelve grade students ($N = 178$) for six forced choice tasks in the aspect 'nature of models' and for another six tasks in the aspect 'purpose of models'. We used the software ACER ConQuest 3 (Wu, Adams, Wilson, & Haldane, 2007) to estimate partial credit models (Masters, 1982) which allow describing the responses to the forced choice tasks on a multi-point scale (Embretson & Reise, 2000). Students' meta-modelling knowledge was inferred using Warm's (1989) Weighted Likelihood Estimator (WLE). We modelled a unidimensional model (general model of meta-modelling knowledge) and the theoretically expected two-dimensional model with the aspects 'nature of models' and 'purpose of models' as two latent dimensions (Krell & Krüger, 2011). The Akaike's Information Criterion (AIC), the Bayesian Information Criterion (BIC) and the sample-size adjusted BIC (ssBIC) were calculated to compare the model-fit (Burnham & Anderson, 2004; Henson, Reise, & Kim, 2007; cf. Krell & Krüger, 2011). The ssBIC index was added, as it was shown to be the most accurate statistic especially for small samples (Henson et al., 2007). All three measures indicate the two-dimensional model to be more appropriate to represent the data (Table 9).

Table 9. Indices for the model-fit of the one- (1D) and the two-dimensional (2D) partial credit models.

Modell	*n*	*m(k)*	*Deviance*	AIC	BIC	ssBIC
1D	178	13	4324	4350	4392	4327
2D	178	28	4083	4139	4228	4089

Consequently, the students' responses to the tasks confirmed the theoretically expected two-dimensionality ('nature of models' and 'purpose of models') on which the proposed test score interpretations are based and

thus provided evidence for validity based on internal structure (AERA et al., 2014). This also means that reliability analyses should be applied for both dimensions individually. This leads to Cronbach's alpha measures of .330 for the aspect 'nature of models' and .704 for the aspect 'purpose of models' (Table 10).[29]

Table 10. Reliability measures for the aspects 'nature of models' and 'purpose of models'.

	$N_{students}$[30]	Grade	N_{tasks}	Cronbach's alpha	EAP/PV (2D)[31]
Nature of Models	178	11-12	6	.330	.399
Purpose of Models	178	11-12	6	.704	.595

The measures of reliability are being discussed in Article 4 with reference to the above mentioned second level of dimensionality which involves the contexts of the tasks as possible further sub dimensions:

Gogolin, S. & Krüger, D. (submitted). Students' understanding of the nature and purpose of models. ***Journal of Research in Science Teaching*****.**

With regard to findings of other researchers which indicate that students' meta-modelling knowledge in biology is context-specific (Al-Balushi, 2011; Krell et al., 2014a, 2012; Lee et al., 2015; Pluta et al., 2011), we investigated the following research question:

[29] Similar values for the aspect 'nature of models' were attained with open-ended tasks by Grünkorn (2014; 'Modellkenntnisse'; 6 tasks; EAP/PV-Reliability .341), with forced choice tasks by Krell (2013; 'nature of models'; 6 tasks; EAP/PV-Reliability .100) and with multiple choice tasks by Terzer (2012; 'nature of models'; 6 tasks; EAP/PV-Reliability .308).

[30] Cronbachs' alpha changes to .662 for the aspect 'purpose of models' when including the students from grade ten (N_{total}=285).

[31] The calculation of the EAP/PV reliability is based on the data for grades eleven and twelve (n = 178) for the two-dimensional model.

RQ Study VII — To what extent does students' meta-modelling knowledge in the aspects 'nature of models' and 'purpose of models' vary across different biological contexts?

285 students from grades ten to twelve completed the questionnaire for the aspect 'purpose of models' (six tasks). As previous research had shown that the tasks for the aspect 'nature of models' were not suitable for grade ten, out of the 285 students, 178 students in grades eleven and twelve also completed the questionnaire for the aspect 'nature of models' (six tasks). We computed the means of students' preferred level of understanding per context und additionally analysed the students' responses to the open-ended justification tasks ($N_{responses} = 720$) by means of a thematic qualitative content analysis (Kuckartz, 2016) where similar students' justifications were grouped together in categories (inductive approach, Kuckartz, 2016; Attachment 109).

The means for students' preferred level of understanding per context showed significant differences within the aspect 'nature of models' but no significant differences across contexts within the aspect 'purpose of models' for grades eleven and twelve. The thematic qualitative content analysis (Kuckartz, 2016) of students' responses in the open-ended justification tasks revealed possible reasons for the differences in students' choice for a level in the forced choice tasks depending on the task context.[32] In the aspect 'nature of models', differences in students' justifications stem primarily from the students' existing knowledge or the status of the knowledge which they attribute to the cargo of the model. For example, for the context of the bio membrane, students most often reasoned that the model is right (level I)

[32] The tables with the categories of the thematic qualitative content analysis can be looked up in Article 4 in chapter 11. The coding scheme for the thematic qualitative content analysis in enclosed in Attachment 10.

because it complies with what they had been taught about the original in school. In the aspect ‘purpose of models’, many students gave reason about why a model cannot be used for purposes on level II or III. For example, for the context of the Jura forest, the students denied the possibility of investigating the growth of plants in the Jura forest because it would need the model to actively change but a model is a snapshot and thus does not change.

4.5.3. Diagnostic procedure

Our aim was to create an efficient diagnostic instrument that gives direct feedback to the teacher or the students themselves concerning their individual meta-modelling knowledge in the aspects ‘nature of models’ and ‘purpose of models’. We decided that, in order to enable teachers to plan individualised interventions, it would be best for them to have one score of the general level of understanding for each of the two aspects per student respectively. Determined by the findings that students’ meta-modelling scores across the present forced choice tasks were context-specific, we need to respect the differences of the single tasks in order to counteract them being an uncontrolled source of measurement error (AERA et al., 2014). So, instead of using the mean of the six scores per aspect, we calculate the general level for each aspect by conducting a multiple regression analysis (forced entry) with the median as the dependent variable (Field, 2013).[33] This procedure enables us to respect possible differences of students’ meta-modelling knowledge across tasks by using them as regression coefficients. The multiple regression analysis results in every student having one measure (predicted value), expressing his or her level of understanding

[33] The diagnostic instrument produces a complete data set as every student answers all tasks and thus we use methods of classical test theory rather than of item response theory.

(level I, II or III) for the aspect ‘nature of models’ and another one for the aspect ‘purpose of models’.

The fit of the regression models to the data were verified by showing that the assumptions of normality of residuals, linearity and homoscedasticity were met. Multicollinearity, which is a thread to the interpretation of the regression analysis as it increases standard errors of the regression coefficients, was checked by three measures: (1) Pearson’s correlation among pairs of variables showed that none of the pairs correlated highly ($r < .33$), (2) the variance inflation factors (VIF) were lower than 10 ($1.02 < VIF_{Nature} < 1.07$; $1.10 < VIF_{Purpose} < 1.36$) for all predictors and (3) the tolerance statistics showed values higher than 0.2 ($.93 < tolerance_{Nature} < .98$; $.73 < tolerance_{Purpose} < .90$) indicating that none of the predictors have a strong linear relationship with other predictors. The values of the Durbin-Watson test are greater than 1 and lower than 3 for both aspects which informs that the assumption of independent errors is tenable. The standardised beta values of the predictors certify that each predictor variable made a significant contribution to predicting the outcome (Table 11).

Table 11. Multiple linear regression analyses for the aspects ‘nature of models’ and ‘purpose of models’.

Nature of models					Purpose of models				
	b	SE(*b*)	ß	*p*		*b*	SE(*b*)	ß	*p*
(Constant)	-.854	.187		.000	(Constant)	-.506	.070		.000
WC	.103	.049	.100	.038	TR	.104	.025	.128	.000
NT	.177	.041	.209	.000	JF	.118	.028	.132	.000
BM	.226	.040	.276	.000	BM	.183	.032	.180	.000
TR	.254	.042	.297	.000	LZ	.263	.030	.295	.000
EV	.250	.038	.323	.000	EV	.262	.028	.313	.000
VS	.380	.057	.329	.000	BR	.300	.031	.314	.000
R^2=.616; Δ R^2=.602					R^2=.766; Δ R^2=.761				

The grade of the students did not make a significant contribution to predicting the outcome in either of the aspects and is consequently, for the time being, not included as a predictor variable. Nevertheless, a growing dataset may allow to include more predictor variables (e.g., grade, school type) to improve the fit of the regression models to the data and subsequently the accuracy of the prediction (Field, 2013). Regarding the fit of the regression models, the values of R^2 indicated that 62 % of the variability in the outcome of the regression analysis for the aspect 'nature of models' and 77 % for the aspect 'purpose of models' are accounted for by the predictors. The small differences between R^2 and adjusted R^2 indicate that the models generalise well. The models which would be derived from the population rather than the sample would account for merely 1.4 % less for the aspect 'nature of models' and 0.5 % less for the aspect 'purpose of models'. Due to the good fit of the multiple regression analysis, we decided to keep working with this algorithm.

In order for teachers to be able to perform a diagnosis in the classroom with an immediate feedback instead of having researchers coming to school or providing them with the statistical analysis, we developed a digitalised version of the final instrument.[34] Teachers can access an online platform (www.userpage.fu-berlin.de/modelle) where they get information about the instrument and the data analysis procedure.[35] The teachers have to follow three steps:

1. Bitte notieren Sie den folgenden Gruppenschlüssel. [e.g., TQz4w]
2. Bitte verteilen Sie den Gruppenschlüssel an die Schüler_innen und führen Sie den Test durch. [http://userpage.fu-berlin.de/modelle/fb]

[34] Christoph van Heteren-Frese was so kind to realise the online adaptation of the questionnaire and I would like to express my gratitude once again at this point.

[35] The link to the online platform has been published in the journal 'Der mathematische und naturwissenschaftlich Unterricht' and in the journal 'Biologie 5-10'. A screenshot of the online platform is in Attachment 11.

3. Schauen Sie sich die Ergebnisse der Befragung Ihrer Schüler_innen an, indem Sie den Gruppenschlüssel hier eingeben. [e.g., TQz4w][36]

The introduction of a group key allows the teacher to specifically access the data of his or her class in the overall data set. Moreover, a repeated diagnosis with the same key allows teachers to evaluate the success of an intervention.

During diagnosis, the students work on a computer and answer six tasks for each aspect. Two global levels of understanding are being automatically computed by means of the multiple regression analysis. The algorithm retrieves each student's score per task from an SQL-database, computes the median as the dependent variable as well as the regression coefficients as the independent variables and generates the predicted value by the following formula:

$$Y_i = (b_0 + b_1X_{i1} + b_2X_{i2} + \ldots + b_nX_n) + \varepsilon_i$$ [37]

The fact that the regression coefficients are being computed de novo on the basis of the whole dataset for every new prediction entails the program to be able to learn and improve the prediction with growing dataset given the questionnaire is being used responsibly by students ideally in a supervised scenario in school.

The results of the regression analyses are being displayed to every student on the last page of the questionnaire by a coloured split-half circle. The left side of the circle standing for his or her level of understanding in the aspect

[36] [1. Please note the following group key. 2. Please distribute the group key to the students and carry out the test. 3. Access the results of the survey with your students by entering the group key here.]

[37] Y is the predicted outcome, b_1 is the coefficient of the first predictor (X_1), b_2 is the coefficient of the second predictor (X_2), b_n is the coefficient of the n^{th} predictor (X_n), and ε_i is the difference between the predicted and the observed value of Y for the i^{th} participant (Field, 2013). The php script for the regression analysis (including explaining remarks) is in Attachment 132.

'nature of models' and the right side for the aspect 'purpose of models'. Level I is denoted by yellow, level II by green and level III by blue. The levels correspond to those described in the 'model of model competence' (Krell et al., 2016). By using the group key, the teacher has access to a table containing all students' levels of understanding (Figure 4).

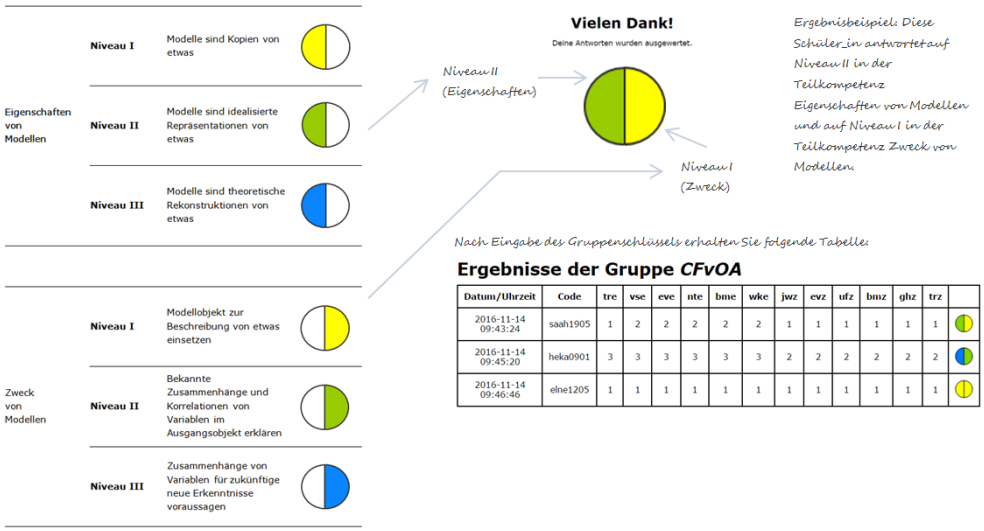


Figure 4. Excerpt from the 'Information for teachers' on the userpage. The figure shows the correspondence of colours and levels for the global levels of understanding computed by the algorithm. For the full 'Information for teachers' pdf see Attachment 133.

With the online tool, the promotion of model competence can either happen immediately after the interrogation by assigning students, depending on their diagnostic result, to differently coloured stations or at a later point after the teacher got an overview of the perspectives present in his or her class. The online version is currently being used by teachers in school.

5. Discussion of Validity Evidence and Consequences of Testing

On the basis of the findings assembled in the last chapters, I now ought to integrate the different pieces of evidence for validity in order to build an overarching validity argument for the use of the final diagnostic instrument. Agreeing with the words of the Standards for Educational and Psychological Testing: "Those making the claims are responsible for evaluation of the claims." (AERA et al., 2014, p. 20), I will weigh the evidence with regard to possible consequences of testing. I will further examine possible distortions in meaning which limit the interpretation of test scores for subgroups of students and I will highlight areas where further research is needed to substantiate the validity argument.

5.1. Test Content

In this subchapter, I will discuss evidence regarding global hypothesis H_{GI} :

- The test content adequately represents the content domain.

Considering the outline of the content domain, the 'model of model competence' (Krell et al., 2016) built a profound theoretical basis for this research project. The decision and the determination to only diagnose students' meta-modelling knowledge in the aspects 'nature of models' and 'purpose of models' helped to narrow down the construct and thus to represent the content domain within a reasonably scoped diagnostic instrument (Hartig et al., 2008). With this restriction in mind, the users of the diagnostic instrument should be careful not to interpret the results any

further than the described construct of meta-modelling knowledge. Terzer (2012), Krell (2013) and Grünkorn (2014) presume that meta-modelling knowledge can be regarded as an indication for model competence but should not be set equal with the latter as model competence per definition involves the practical skills and the willingness to work with and think about models (Upmeier zu Belzen & Krüger, 2010).

Concerning the adequate representation of the content domain, Messick (1995) proposes 'construct-irrelevant variance' and 'construct underrepresentation' as the two major threats to H_{GI} 1. These threats ought to be discussed in the following paragraphs. With reference to construct-irrelevant variance, Messick (1995) highlights the following:

> „The concept of construct-irrelevant variance is important in all educational and psychological measurement, including performance assessments. This is especially true of richly contextualized assessments […]. And it matters whether the contextual clues that people respond to are construct-relevant or represent construct-irrelevant difficulty or easiness." (p. 743)

In order to avoid construct-irrelevant variance, we decided to include as little information as possible in the task stem. We merely referred to the following pictures and pointed out that the model was made by scientists (as this is considered construct-relevant information) but we did not give any further content-related information. By doing so, we differed from Terzer (2012), Krell (2013) and Grünkorn (2014) who provide essential content-related information so to create a knowledge basis for students. Although, in the empirical studies, we did not encounter evidence that would lead us to think that the lack of content information in the task stem posed a problem for the students to express their meta-modelling knowledge, the missing information could have disadvantaged some students due to their lack of prior knowledge limiting their access to the task. This might account for some of the observed unfairness in testing for the subgroup of younger

students. A recent study by Ropohl, Walpuski, and Sumfleth (2015) investigated into advantages and disadvantages of chemistry competence assessment tasks which either did or did not include a task stem with content-related information. The researchers conclude that tasks including a task stem with content information are more likely to assess competences in the sense of Weinert (2001) while tasks without a task stem are more likely to assess declarative knowledge. Ropohl et al. (2015) also accredit higher test fairness to the tasks including task stems with content information and a better coverage of the lower ability spectrum. Thus, although we tried to avoid construct-irrelevant variance by information in the task stem, we may have caused unfairness for some students by not balancing the variance in prior knowledge. The issue of how much information should be included into the task stem remains unclear and requires further systematic investigation.

Regarding 'construct underrepresentation', which is, according to Messick (1995), the other major threat to H_{GI} 1, the experts confirmed that the test content is relevant to the construct under investigation and that the created answer options map the theoretical descriptions. This produces evidence against construct underrepresentation and, hence, H_{GI} 1 could not be refuted by the expert judgement. The discussion with the experts helped to discard unfitting answer options and to improve other answer options with regard to possible difficulty generating task characteristics (Kauertz, 2008). On a more global level, the experts were also asked to judge the model contexts with regard to their representation of the concrete-abstract continuum (Krell et al., 2014b) which resulted in the construction of two more contexts to sharpen the 'abstract' end. Although empirical studies served as a basis for the development of the answer options (Grünkorn, 2014) and the choice of model representations for the contexts (Krell et al., 2014b), it would have been more precise to, in a first step, have students answer open-ended tasks with regard to the newly designed contexts rather than using the existing

statements from the work of Grünkorn (2014). Same is true for the representational form of the newly chosen models on the concrete-abstract continuum. The classification should have been double checked again empirically with students. This goes in hand with the repeatedly expressed claim by Messick (1995) that expert judgement is fallible and must be supported by other evidence.

5.2. Response Processes

Borsboom and Markus (2013) summarise that "Somewhere in the chain of events that occurs between item administration and item response, the measured attribute must play a causal role in determining what value the measurements outcomes will take; otherwise, the test cannot be valid for measuring the attribute."

H_{GI} 2 operationalises this request for evidence based on students' response processes:

- The expected performances measured by the test are actually performed by students when taking the test.

In our test development project, H_{GI} 2 would be supported by students who choose answer options for logically sound reasons that support their choices. On the other hand, H_{GI} 2 would be threatened by students, who choose an answer option on a certain level because they misinterpret the level that was being operationalised by the chosen answer option, or by students, who use test-taking strategies to pick an answer like 'Elimination' and 'Educated guessing' (Famularo, 2007; Leighton, 2004; Treagust et al., 2004).
Concerning the first threat, there are two possible misinterpretations of an answer option. A student either interprets the answer option into a higher or

into a lower level. Misinterpretations in direction of a lower level are a more severe thread to validity because it would ultimately result in an overrating of a student by the diagnosis and thus in more disadvantageous unintended consequences of testing (AERA et al., 2014; Shepard, 1997). The proposed interpretation of scores could in these cases not be supported by the empirical evidence (AERA et al., 2014). This outcome is problematic in regard to the interpretation of the results as a teacher might choose a promotion task that is not adapted to the students' perspective. The empirical approaches (studies IV, VI, VII) taken in this research project to check for misinterpretations generated fruitful data that allowed to check for level-specific, context-specific and grade-specific misinterpretations. It ultimately helped to select answer options and contexts for the tasks (Gogolin & Krüger, 2016b) and to exclude subgroups of students for whom the diagnostic tasks will not generate validly interpretable results (Gogolin & Krüger, submitted, 2016a).

In the course of task development, we had to dismiss some of the constructed answer options because the fit between interpreted and theoretically intended level decreased with higher level in both aspects (study IV: 'nature of models' and study VI: 'purpose of models'; Table 3). In Article 3, we reason that, if the students had a naïve understanding of the nature or the purpose of models (Grosslight et al., 1991; Grünkorn, 2014), they may not agree with the level III answer option, and, further, they would not be able to appreciate the epistemological notion inherent in the level III answer options. Alternatively, the students could be reinterpreting the answer options in direction of their own perspective on models. This strategy had been described before by Krell, Czeskleba, and Krüger (2012) as an observation during a think aloud study. We also had to dismiss some of the contexts entirely. It was not possible to find appropriately understood answer options for the contexts 'Air stream' and 'Water melon' in the aspect 'nature of models' and for the context 'Air stream' and 'Bacterial

growth' in the aspect 'purpose of models'. In Article 3, we hypothesised that the representational form of the models in these contexts may have caused some of the misinterpretations. Looking at the concrete-abstract continuum (Krell et al., 2014b) that was used to choose and classify the models for the contexts, the four aforementioned models are the most abstract because all of them are strongly formalised. Cohors-Fresenborg et al. (2004) examined PISA tasks and concluded that formalisation was one of the most difficulty generating task characteristics. Hence, we suspect the representational form of these models to impede the students from uttering their perspectives regarding the corresponding tasks. The contexts had to be discarded. After all, the repeated test with the final forced choice tasks (open-ended justification tasks; study VII) showed that the answer options were being understood as intended by the proposed sample of students. This observation supports H_{GI} 2 and creates evidence that the students' future scores will represent their meta-modelling knowledge validly.

The think aloud protocols, the student interviews and the open-ended justification tasks contributed successfully to gathering evidence based on students' response processes. This kind of evidence is often omitted in the validation process as it is a strenuous and time consuming process (Adams & Wieman, 2011; Leighton & Gierl, 2007b). Leighton (2004) also argues that "one of the stumbling blocks" (p. 6) to using data from verbal reports are negative prejudices concerning its accuracy and trustworthiness. This is why I would like to include a short methodical discussion and judge the quality and the usefulness of the data gained by the different approaches in this project.

In study IV, we conducted both 'concurrent think aloud protocols' and 'retrospective descriptions of thinking'. In the concurrent think aloud protocols, we had a very limited number of student responses as we analysed only those parts of the protocols in which the students referred to their chosen answer option. In order to enlarge the sample size for the

selection of answer options for the final forced choice tasks, we merged the students' responses from the 'concurrent think aloud protocols' and from the 'retrospective descriptions of thinking'. It has to be highlighted here though, that the findings of both approaches should not imprudently be put on the same level. Although the descriptions of thinking are often referred to as 'retrospective thinking aloud' (Ericsson & Simon, 1998; Sandmann, 2014), Ericsson and Simon (1998) point out that the results are not identical with concurrent think aloud protocols because the retrospection triggers an explanation of thinking and therefore may change the sequence of thought. Ericsson and Simon (1998) add that this kind of triggering usually improves the answer by the student. This often criticised circumstance (e.g., Nisbett & Wilson, 1977) may not be considered problematic for the present project. With neither of the two approaches, we were interested in reconstructing the authentic problem-solving process but in capturing how the students understand our answer options. The interpretation of the students' responses was the same for both approaches, although, theoretically, with the 'concurrent think aloud protocols', we could have obtained more information concerning the problem-solving process. Considering, that our forced choice tasks do not contain a problem, but ask students to compare their perspective with the given ones, I retrospectively come to the conclusion that concurrent thinking aloud is not the most productive approach for our tasks as its potential in eliciting the sequence of thought is not being used. A systematic approach with retrospective descriptions of thinking as performed for the aspect 'purpose of models' (study VI) or with open-ended justification tasks as performed for the final tasks (study VII) would have given us the same information with less effort and thus would have provided us with a better data basis for the selection of answer options for the aspect 'nature of models'.

5.3. Relations to Other Variables

Evidence based on relations to other variables shall be discussed hereinafter with reference to H_{GI} 3:

- Theoretically expected patterns of relationships of the test under investigation with external variables can be identified in students' test scores.

Data from the monotrait-multimethod approach (study III) may be used to draw on convergent validity. The conversion of measurements in study III for the students from grades eleven and twelve supports H_{GI} 3 and therefore strengthens the argument that the interpretation of the scores representing the underlying construct is valid. Still, many researchers point out that the interpretation of evidence based on relations to other variables should be done with great care due to possible method-specific influences (Campbell & Fiske, 1959; Chaoui, 2011; Martinez, 1999; Moosbrugger & Kelava, 2012). The results of study III were discussed with regard to validity evidence in Article 2 (Gogolin & Krüger, 2016a). While in Article 2, we mainly try to gather explanations for why there was no evidence for convergent validity for students from lower grades, I will now scrutinise the power of the evidence we have for convergent validity for grades eleven and twelve.

In grades eleven and twelve, the students did not answer significantly different across the three different measures, thereby indicating convergent validity. The three-dimensional PCM was assumed to fit the data appropriately and support the soundness of the validity evidence. On the other side, we found low person reliabilities (EAP/PV) for all of the assessment methods (forced choice tasks: 0.508, open-ended tasks: 0.550,

interview questions: 0.538). There are a number of possible reasons for the low person reliabilities within the forced choice tasks which are to be discussed in the next chapter (chapter 5.4). Concerning the open-ended tasks and the interviews, we ought to recall that reliability values depend on the number of tasks on a scale (Cortina, 1993; Crocker & Algina, 1986; Field, 2013) with the more tasks the better. Since we only obtained two data points from each of the open-ended diagnostic tasks and the interview questions, we cannot expect good values of reliability for these measures (Field, 2013). Although the PCM seemed to fit the data, the low reliabilities threaten the consistency of the measures and thereby weaken the trust in the claim made in H_{GI} 3. Additionally, it has to be said, that we did not use the final version of the forced choice tasks in this study, so, ideally, the study should be repeated with the final tasks (Raykov & Marcoulides, 2011).

The intervention study provided an opportunity to gather further evidence for validity based on relations to other variables. It was expected theoretically that students would have better test scores after an intervention than before. The data gathered before and after the intervention reflected this relationship as predicted. The findings might be interpreted as supporting H_{GI} 3, but caution is to be advised with regard to this power of this evidence. It has to be noted here, that although I avoided any phrasing from the instrument during the intervention, it cannot be ruled out that the students' post-instruction answers to the tasks were in some part influenced by the language used during the intervention ('teaching to the test'; Moosbrugger & Kelava, 2012). Shepard (1997) stresses the relevance of this problem which was originally raised by Gulliksen (1950):

> „In the case of coaching or teaching to the test, the threat to validity is not just that use of test results does not have the intended effect on learning; it is also that a flaw in the conceptualization of the test made it susceptible to invalid score gains that then render its use invalid." (Shepard, 1997, p. 7)

The crux of the matter is that if 'teaching to the test' can easily improve scores in the test, without improving scores in other measures set to measure the same construct, then the validity of the test is rather limited. Considering our test consists of short phrases that include operators set to define a level of understanding models, then that threat is likely to have an impact on the validity of the score interpretation. We did not perform a follow-up investigation which may have reduced the likeliness of an influence of the language due to the timely distance. Nevertheless, more evidence can be produced in the future on the basis of pre- and post-data from further interventions by teachers.

Summarizing, we did not gather sufficient evidence based on relations to other variables in this research project. H_{GI} 3 cannot be refuted but the evidence to support it needs to be strengthened by further investigations. The empirical investigation of the predictive power of the diagnostic instrument has so far not been part of this research project and thus, long reaching consequences of testing with the instrument (McCoach et al., 2013) cannot be judged. We decided to exclude this kind of evidence for now because it is not central to our argument (AERA et al., 2014; Kane, 2015). The instrument ought to be used to assign students to focussed and individualised learning groups rather than to make selective decisions about students' grades or their future. Then again, once the instrument is used in school in combination with interventions, we could gather some data on the grades of the students at the end of the year and possibly, there will be a correlation between the students' score after the intervention and their grade in biology. This strongly depends on the work of the teacher and his or her focus in the exams though. Furthermore, in order to pay credit to the contemporary notion of validity, which is concerned with test score interpretation rather than test scores (MacIver, Anderson, Costa, & Evers, 2014), we would foremost be interested in how teachers use the instrument

and which consequences of testing with the instrument they derive for their students.

5.4. Internal Structure

In the present chapter, I ought to discuss to what extent evidence for validity based on internal structure could be generated during the research project. The evidence was operationalised in H_{GI} 4:

- Theoretically expected patterns of relationships among test components can be identified in students' test scores.

As mentioned in chapter 2.2., evidence for validity based on internal structure includes the aspect 'Fairness in Testing for Subgroups'. Relating this aspect to H_{GI} 4 seems far-fetched at first, but it can be realised by theoretically not expecting differences across subgroups of test takers and empirically searching for differences (AERA et al., 2014; Kunnan, 2007). During this research project, it was shown empirically on numerous occasions (studies III, IV, VI and VII) that there are grade-specific differences in students' understanding of the developed answer options for the forced choice tasks in the aspect 'nature of models' and 'purpose of models' respectively. Considering the findings of these studies, we need to be careful when wishing to use the tasks in grades seven to ten because the inferences that will be drawn from the scores for this subgroup are less well supported by empirical evidence for validity (AERA et al., 2014). In terms of Messick (1995), this outcome allows to suspect aspects of the test to produce construct-irrelevant variance (e.g., inappropriate sampling of test content, disparities in test context, variety of test responses or opportunity to learn).

The final forced choice tasks in the aspect ‘nature of models’ are therefore only recommended for grades eleven and twelve and the tasks for the aspect ‘purpose of models’ for grades ten to twelve. In the following section, I will use arguments brought forward in Article 2 to discuss possible reasons for why a valid score interpretation cannot be supported for students in grades seven to nine (respectively ten). I will then also evaluate consequences of testing with regard to the use of the final instrument in lower grades.

The analysis of the students’ interpretations of the answer options revealed that in both aspects, the students most often misinterpreted the answer options referring to level III of the ‘model of model competence’ (Krell et al., 2016). The students had trouble understanding the notion of a model being a theoretical reconstruction (‘nature of models’) which can be used to derive hypotheses about an original (‘purpose of models’). Students in our studies frequently confused the hypothetical nature attributed to a model on level III with a general or personal lack of information about the original.

> “Man hat über Membranen schon sehr viel heraus[gefunden], aber ich denke nicht, dass man schon alles weiß, um sie zu 100% richtig darzustellen. Ich denke jedoch, dass Wissenschaftler mit den Informationen, die sie haben, die Membran richtig darstellen.“
> „Ich weiß nicht, ob es auf dieser Welt noch andere Grippeviren gibt, die anders aussehen könnten, deshalb habe ich vorsichtshalber diese Antwort angekreuzt.“[38]

This lead us to argue that students in lower grades might not have a profound understanding of what defines a scientific hypothesis and they rather see them as “(educated) guesses” (Gibbs & Lawson, 1992; McComas, 2013). In contrast to this, the American College Dictionary (Barnhart, 1953) defines the word hypothesis as “a proposition (or set of propositions)

[38] “There have been a lot of studies on membranes, but I do not think everything is known in order to represent it 100% correctly. I think, however, that scientists, with the information they have, represent the membrane just right.”
"I do not know if there are other influenza viruses in the world that might look different, so I checked this answer as a precaution.”

proposed as an explanation for the occurrence of some specified group of phenomena" (p. 596). Moreover, the students seem not to be aware that a scientific investigation involves the formulation and testing of hypotheses. They expressed not to see a point in deriving hypotheses in order to investigate a phenomenon.

> „Ich habe mich für die erste Antwort entschieden, da ich denke, dass ein Modell in erster Linie etwas veranschaulichen soll. Man kann es nicht voraussagen, wie sich der Schädel weiter entwickelt, da dies nur auf Vermutungen beruht."[39]

Reasons for this disarray can only be speculated about. Many researchers have argued that teachers have a rather limited understanding of the nature of models and use models primarily to demonstrate established knowledge (Borrmann et al., 2014; Crawford & Cullin, 2004; van Driel & Verloop, 2002); a practice which would shape students' perspectives towards models in direction of level I and II in both aspects. Meisert (2007) describes the observation that hypotheses are used by teachers to integrate students' preconceptions into the lessons. This practice would lead to hypotheses not being seen as quoted above but as mere "guesses" which ought to be corrected or verified in the course of the lesson. Gibbs and Lawson (1992) underline these points and link students' uneducated understanding of the term 'hypothesis' to the way it is employed in textbooks: "hypotheses are not merely educated guesses based upon collected information as some textbook authors, and perhaps most high school teachers, would lead you to believe" (p. 139).

Independent from the just mentioned reasons, the obvious misinterpretations may be seen as an indication for students in lower grades not to consider models as theoretical reconstructions or as tools for scientific investigations.

[39] "I chose the first answer, because I think that a model is primarily intended to illustrate something. One cannot predict how the skull will further develop, since this is based solely on assumptions."

Under the perspective of validity, this conclusion shall be used as an argument for why the final diagnostic instrument is not intended or suitable for the subgroup of students from grades seven to nine. If none of the students held perspectives on level III, than all of the students in this subgroup would need to be fostered and the instrument would not have as much diagnostic value. Consequently, for students from grades seven to nine, I recommend a general promotion. Teachers are in this case better advised to use the open-ended diagnostic tasks by Grünkorn (2014) if they still wish to distinguish perspectives on level I and II (cf. Grünkorn, Lotz et al., 2014).

With regard to the consequences of testing, the Standards for Educational and Psychological Testing (AERA et al., 2014) bring forward the point that the strictness of the interpretation of validity evidence depends on the proposed uses of the test. The strictness increases as the consequences of decisions and interpretations grow in importance. As the results of the analysis were so far, in research articles, communicated to a great audience, we decided that a higher degree of strictness is warranted. In the classroom, the diagnostic decisions attained with the instrument can be corroborated by information from other sources and the initial decision can be easily corrected. Thus, the diagnostic instrument, when carefully interpreted, may also be used in grade nine classrooms.

Another aspect of evidence for validity based on internal structure is the dimensionality of the construct to be measured. This question has been intensively investigated and discussed in the literature (Grünkorn, 2014; Krell, 2013; Terzer, 2012) because it has consequences for the diagnosis of students' meta-modelling knowledge as well as for its promotion. The fundamental question of whether meta-modelling knowledge is global or aspect-dependent has been investigated by Krell et al. (2014c) who conclude that "students seem to have a complex and at least partly inconsistent pattern of understanding models" (p. 1). Focusing the attention solely on the

'model of model competence' (Krell et al., 2016), there are three possible structures of meta-modelling knowledge to be extracted. The one-dimensional model, comprising all aspects as one (Terzer, 2012); the two-dimensional model, distinguishing between "ontological and epistemological concepts of models" on the one side and "cognitive processes while reflecting the act of their use in science" on the other side (Grünkorn et al., 2014, p. 6); and a five-dimensional model with each of the aspects as one dimension (Krell et al., 2016). The analysis of dimensionality in the present project points at either the two-dimensionality or the five-dimensionality of meta-modelling knowledge. The dimensionality cannot be entangled further with the present data because we only included two of the aspects which each also belong to a different dimension within the two-dimensional model. Being primarily concerned with the validity of the proposed score interpretations from the diagnostic instrument rather than with the validity of the 'model of model competence', we aimed at investigating whether the a priori expected two-dimensionality was being reflected in the data. Since all indices (AIC, BIC, ssBIC) inform that the two-dimensional model is the most appropriate to represent the data, we gathered evidence that supports H_{GI} 4 and therewith a valid score interpretation. From a didactical point of view, the difference between the aspects 'nature of models' and 'purpose of models' is beneficial for the interpretation of the diagnostic instrument of this research project because it allows for a more differentiated promotion of meta-modelling knowledge (Fleige et al., 2012; Fleischer et al., 2013; Krell & Krüger, 2011). Interestingly, although the two aspects could be distinguished statistically and the students seem to adopt perspectives on different levels across the aspects, there were some obvious interactions between the aspects in the qualitative data. Students often explained their answers in the aspect 'purpose of models' with reference to what they believe the nature of a model should be, for example:

> „An dem Modell wird nicht weiter geforscht, da ein Modell nur das zeigen kann, was schon erforscht ist.“[40]

According to Boumans (1999), model building and model justification must be logically intertwined. Boumans (1999) analysed models presented in research papers and suggests that the artificial separations of model building and model justification is a possible reason for why models might be seen as end products of science:

> “In each case, these characteristics were presented at the end of their paper, giving the strong suggestion that the justification of their model was disconnected from the building process. This is how it was presented in each publication. Such presentation hides the actual process of model building, which is more like a trial and error process till all the ingredients, including the empirical facts, are integrated. In other words, justification came not afterwards but was built-in.“ (p. 95)

The perceived relationship between a model and the referring original might be a profound influence when thinking about models in general. Capturing or defining the nature of the relationship between a model and a phenomenon has been debated about extensively in the philosophy of science (Bailer-Jones, 2009; Giere, 2001; Magnani et al., 1999; Morgan & Morrison, 1999). When considering models as entities that describe empirical reality (Bailer-Jones, 2009), one ought to ask how models do that and to what degree. Giere (2001) indirectly supports Boumans’ (1999) claim of the interconnection of model building and model justification when he described that the correspondence between a model and the real world objects cannot be judged directly or certainly. Models can serve as hypotheses that can be compared with data to judge the fit of the model to the ‘real world’ (Giere, 2001).

[40] “The model is no subject to research, since a model can only show what has already been explored.”

Considering models as theoretical reconstructions should therefore facilitate perspectives on level III in the other aspects because no answers about the 'real world' might be found without the process of hypothesis derivation, testing and changing. On the other side, not considering models as such but seeing them solely as products of science might hinder the students from recognising perspectives on level III in the other aspects. Terzer (2012) highlights in her work that the aspect 'nature of models' is rather strongly connected to all other aspects ($.591 < r < .682$; Terzer, 2012). Grünkorn (2014) discusses whether the parallelisation between original and model is a kind of basic skill that affects all other aspects. Hence, the relationship between a model and the referring original might be seen as a basis on which all other considerations about models rely.

The question of dimensionality was broadened in the present work with regard to the issue of reliability. We theoretically expected the scales within the aspects 'nature of models' and 'purpose of models' to be unidimensional. So, in order to support H_{GI} 4, we would need to find good or excellent measures of reliability among the tasks in each aspect (Field, 2013) because reliability values measure the strength of the unidimensionality of a scale (Field, 2013). We did not produce such values of Cronbach's alpha but rather values between .33 ('nature of models') and .70 ('purpose of models'). The discussion of why we have such low values brings up two different grand strands of argumentation. Either, we were right in theoretically assuming unidimensionality of each of the two scales and the low reliabilities are due to construct-irrelevant variance resulting from flaws in the task development process. Or, we did not make the right theoretical assumption in the first place and the scales for the two aspects are not in themselves unidimensional. A context-specificity of students' meta-modelling knowledge would fragment each of the scales and make good reliability values impossible. Cronbach (1951) suggests that reliability should be calculated separately for the tasks relating to different sub

dimensions (meaning in our case, for example, sub dimensions within the aspect 'nature of models' for which we might not have more than one task each). In this case, we would face the problem of construct-underrepresentation because we wouldn't have enough tasks. In either of the two scenarios we violate the assumption of validity. A quote by Cronbach and Meehl (1955) picks up on consequences that derive from this problem:

> "[The test developer] is free to say to himself privately, 'If my test disagrees with the theory, so much the worse for the theory.' This way lies delusion, unless he continues his research using a better theory". (p. 296)

Evidence by other researchers (Al-Balushi, 2011; Krell et al., 2014a, 2012; Lee et al., 2015; Pluta et al., 2011) as well as our own data from the open-ended justification tasks (study VII) strongly point to the conclusion that students' meta-modelling knowledge is context-specific.

Science philosophers and science education researchers (Bailer-Jones, 2002; Giere, 2001; Mahr, 2008, 2011; Odenbaugh, 2005; van der Valk et al., 2007) define models only in the context of their use and thereby allow all possible perspectives on models. In a study with scientists, Schwartz and Lederman (2005) state that: "Seventeen of the 24 (70.7 %) scientists indicated models were explanations or ways to organise observations that also involved testing predictions." (p. 7). When philosophers and scientists have a wide range of perspectives that they apply dependent on the model context, why should students not, in a slightly flattened manner, also hold numerous perspectives that they apply in function of the context? Guerra-Ramos (2012) raises a similar point concerning concepts about nature of science and argues that different ideas can be applied in different situations.

The differences in students' justifications that occurred in the analysis of the open-ended justification tasks in study VII need to be interpreted with care because the students justified their responses in the forced choice tasks with a clear reference to these tasks. Hence, they were likely to be influenced by

the answer options given in the tasks. Although the students' justifications went further than what was being said in the answer options, we cannot clearly determine whether the students' genuine meta-modelling knowledge was context-specific.

If students' meta-modelling knowledge was indeed context-specific, it would make it necessary to include further tasks for each possible sub dimension in order to decrease construct-underrepresentation and increase reliability. This generates two challenges: (1) the context-specific sub dimensions need to be thoroughly theoretically described and (2) the length of the diagnostic instrument would grow immensely. An enlargement of the instrument was to be avoided though because it would most likely lead to losses in students' concentration and willingness to solve the tasks carefully (Optimizing-Satisficing Problem; Krosnick, 1999; Moosbrugger & Kelava, 2012).

> „Bei der Festlegung der Testlänge darf die Praktikabilität des Tests und insbesondere die Motivationslage der Probanden nicht aus den Augen verloren werden. Je länger der Test, desto mehr ist damit zu rechnen, dass die Items nicht mehr konstruktgemäß bearbeitet werden. […] Das Optimizing-Satisficing-Modell (Krosnick, 1999) …] macht deutlich, dass bei einem von den Testteilnehmern subjektiv als zu lang empfundenem Test die Bearbeitungsqualität sinkt, sodass nicht mehr von einem adäquaten Testergebnis ausgegangen werden kann." (Moosbrugger & Kelava, 2012, p. 35)[41]

Also, and this was one of the fundamental pillars of this research project, the diagnosis with our instrument was to be efficient and thus as short as possible.[42] The dependence of Cronbach's alpha on the number of questions

[41] "When determining the test length, the practicability of the test and, in particular, the motivation of the test takers must not be lost sight of. The longer the test, the more can be expected that items will not be answered appropriately. [... The Optimizing Satisficing Model (Krosnick, 1999) ...] makes clear that the answer quality for a test, which is subjectively perceived as too long by the test participants, decreases, so that an adequate test result can no longer be assumed."

[42] A further in depth discussion of this dilemma has been performed in Article 4.

and the variance in the scores has proved to be a problematic point in a number of biology education studies (Schmiemann, 2010; Terzer, 2012; Wellnitz, 2012). It became clear in our analysis that interpreting Cronbach's alpha can be difficult, because it has to be verified whether score differences are a consequence of real differences within the assessed construct (e.g., multiple factors) or if they are due to a measurement error (Cronbach & Shavelson, 2004). For the purpose of validating our score interpretation with regard to reliability, it should be more instructive to perform a retest-reliability study (Adams & Wieman, 2011; Field, 2013). Test-retest coefficients would be obtained by administering the same form of the test on two distinct occasions and computing Pearson's product-moment correlation coefficient (r) or the Intraclass Correlation Coefficient (ICC) (Field, 2013). We clearly need more investigations into the context-specificity of students' meta-modelling knowledge using methods that do not confound validation and investigation purposes.

Students' context-specific responses, whether they might be due to the construct or the task characteristics, were respected in the diagnostic procedure by means of multiple regression analyses with the tasks as predictor variables (Field, 2013). The grade of the students, as another theoretically expectable independent variable (Al-Balushi, 2011; Chittleborough et al., 2005; Grosslight et al., 1991; Lee et al., 2015) was not included in the regression because it turned out not to improve the prediction of the outcome (students' assigned meta-modelling knowledge) significantly.

Summarizing, we were able to provide some evidence for validity based on internal structure (H_{GI} 4) by empirically reproducing the theoretically expected two-dimensionality of students' meta-modelling knowledge (Krell & Krüger, 2011). With regard to the context-specificity of students' meta-modelling knowledge, I cannot make a final statement but there are indications that students' meta-modelling knowledge is context-specific (Al-

Balushi, 2011; Krell et al., 2014a, 2012). This is a possible explanation for the low reliabilities in all studies using contextualised tasks to diagnose students' meta-modelling knowledge. Context-specificity would have consequences both for assessment and teaching (Krell et al., 2014a; Lee et al., 2015; Schwarz, 2002).

6. Diagnosis and Promotion

As early as in the introduction, I raised the question 'What does it help to know how students understand models?'. At the end of the day, it helps foster students successfully (Adams & Wieman, 2011; Günther et al., 2016; Oh & Oh, 2011). Individual diagnosis allows teachers to select appropriate interventions for individual students. For this conjunction to be feasible, the diagnostic result needs to be simple and easily accessible (Adams & Wieman, 2011). Furthermore, the teacher needs to dispose of promotional measures which he or she can apply in case of a certain diagnostic result (Günther et al., 2016; Oh & Oh, 2011). The combination of both constitutes a sine qua non (Hartig et al., 2008). In the two subchapters of this section, I will first summarise diagnostic results for students' meta-modelling knowledge which might help to get an orientation as to what a likely distribution in grades eleven and twelve in German secondary schools may be. I will then proceed to summarise aspects that might be helpful for the promotion of students' model competence and demonstrate some concrete ideas for the biology classroom.

6.1. Students' Meta-modelling Knowledge

After having collected a somewhat reassuring amount of evidence for a valid score interpretation, we used the paper-pencil version of the final diagnostic instrument to diagnose students' meta-modelling knowledge. Article 4 reports design and findings of this study VII.

Gogolin, S. & Krüger, D. (submitted). Students' understanding of the nature and purpose of models. *Journal of Research in Science Teaching*.

In Article 4, we analysed students' meta-modelling knowledge with regard to context- and grade-specific differences. As these analyses have been referred to previously under the perspective of validity, I will now focus solely on describing and discussing the final diagnostic results with reference to the theoretical basis of the 'model of model competence' (Krell et al., 2016).

$RQ_{\text{Study VII}}$ How frequently are the three levels of understanding of the nature of models and the purpose of models represented among students?

In chapter 3.2., I pointed out that students' meta-modelling knowledge of the nature and the purpose of models has been described differently in empirical studies. Against the background of the gathered validation evidence, it becomes all the more important to constrict the given empirical basis to studies directly comparable to the present research project.
For the aspect 'nature of models', the work of Treagust et al. (2004) as well as Chittleborough et al. (2005) lead to hypothesis $H_{\text{GII Nature}}$

- The majority of students show a meta-modelling knowledge of the nature of models on level III.

For the aspect 'purpose of models', the works by Krell (2013) was used to derive hypothesis $H_{\text{GII Purpose}}$

– The majority of students show a meta-modelling knowledge of the purpose of models on level III.

In order to produce data to answer the research question, we had students from grades eleven and twelve who answered the tasks for the aspects 'nature of models' ($N_{students}$ = 178) and students from grades ten to twelve who answered the tasks in the aspect 'purpose of models' ($N_{students}$ = 285). The students' responses were analysed by means of the multiple regression analysis and one, respectively two, global levels of meta-modelling knowledge (one for the aspect 'nature of models' and one for the aspect 'purpose of models') were assigned to each student.

In the aspect 'nature of models', the majority of students chose perspectives on level II indicating that models are idealised representations of their corresponding original. In the aspect 'purpose of models', students' were primarily attested a meta-modelling knowledge on level I indicating that they see the purpose of models in describing the corresponding original.

To sum up, the data in the aspect 'nature of models' does not support $H_{GII\ Nature}$ because the findings indicate that students in the present study see models mainly as representations of the original (level II) to which they relate rather than as representations of ideas of how things work. The observations made in the present study align more with the studies of Grosslight et al. (1991), Harrison and Treagust (1996), and Sins et al. (2009). Grosslight et al. (1991) state that the interviewed eleventh grade students were most likely to think of models as representations of reality rather than as constructions that embody different theoretical perspectives. Sins et al. (2009) likewise report that the majority of students were assigned to level II (65 %), followed by level I (23 %) and level III (12 %).

Concerning the aspect 'purpose of models', the present findings do not support $H_{GII\ Purpose}$ because the majority of students showed a meta-modelling knowledge on level I and II rather than as predicted on level III.

Nevertheless, the present findings are similar to the findings of a number of other studies (Chittleborough et al., 2005; Grosslight et al., 1991; Harrison & Treagust, 1996; Treagust et al., 2004).

Before elaborating on possible explanations for the present findings, I would like to remark that our findings, as well as those of other researchers, are limited in their informative value. Influences that impact empirical findings across studies could be, among other things, divergent theoretical frameworks, different methodical approaches or differences across the groups of students that were investigated. Furthermore, the influence of a context-dependency is highly likely and would in return lead to divergent findings depending on whether or not the tasks were contextualised and if so, which contexts had been employed. As a consequence of these considerations, the findings of the present study can and should not be generalised.

Reasons for the presently observed meta-modelling knowledge on level I and II among students of eleventh and twelfth grade are manifold. The perspective on models as replications of reality (level I in the aspect 'nature of models') might stem from the students being ignorant of how a model is built from data. They seem not to consider that scientific knowledge (in all its forms) is "subjective (involves personal background, biases, and/or is theory laden), necessarily involves human inference, imagination and creativity (involves the invention of explanations), and is socially and culturally embedded." (Lederman & Lederman, 2014, p. 601). A shortage of understanding the nature of science and the role models play in the process of scientific inquiry might have an influence on students' meta-modelling knowledge. Gobert et al. (2011) articulate the importance of a connection between understanding models and understanding the nature of science:

> "The understanding of scientific models is an important component of students' understandings of the nature of science as a whole .... the key

> connection between the nature of models and the nature of science relates to the belief that models are to be viewed as not completely accurate from a scientific point of view; that is, they are tentative and open to further revision and development." (p. 657)

The idea that models might be a 'door opener' to a better nature of science understanding was put forward by a number of researchers (Gobert et al., 2011; Leisner-Bodenthin, 2006; Passmore et al., 2014). One might argue in return though, that understanding the nature of science is an important component of students' understanding of the nature of models. The students in our study most likely did not have a profound understanding of nature of science because they indicated that scientific knowledge is not being gained by using models. Furthermore, the students ignored the importance of hypotheses for conducting investigations. These rather naïve ideas about science in general might hinder students from adopting perspectives on level III in the case of models and modelling. This argument is supported by Chi (1992) who discriminates between conceptual change that occurs within an ontological category and one that necessitates a change between ontological categories. We suppose that school and science are two distinct ontological categories for students. Trier et al. (2014) found in their interview study that some students showed a more elaborate meta-modelling knowledge when talking about models and modelling in the context of science than in the context of school. For students to understand models as methods of science, they cannot simply borrow predicates and properties of the school category to interpret events in the science category. Adapting Chi's (1992) criteria for conceptual change across ontological categories for the context of models in science, students would, in reverse to common suggestions (Gobert et al., 2011; Leisner-Bodenthin, 2006; Passmore et al., 2014), first need to learn about nature of science, since a prerequisite for crossing ontological barrier is to have some understanding and knowledge of the target category to which the concept is to be changed.

This approach might be just as arduous though because teaching students' about nature of science seems to be in no way easier than teaching about the nature of models. Höttecke and Rieß (2015) explain the dilemma with reference to the topic of experiments (their statements might be, without adaptation, transferred onto the issue of models):

> "Lehrkräfte durchdenken das Thema Experimentieren aus der Perspektive der Vermittlung von Fachwissen (Flick, 2000). […] Eine klare Orientierung am Lehren und Lernen über die Natur der Naturwissenschaften ist kaum zu erkennen, weil es an didaktischen Strategien z. B. zum Umgang mit Offenheit und Unsicherheit mangelt (Ruhrig & Höttecke, 2015). Es verwundert daher nicht, dass auch Schüler/innen einen klaren Unterschied machen zwischen Experimentieren im Unterricht und Experimentieren in den Naturwissenschaften (Henke & Höttecke, 2013). Wenn das eigene Experimentieren der Schüler/innen als Erfahrungsressource dienen soll, um auf Experimentieren in der Forschung reflektieren zu können, dann müssten Lerngelegenheiten Reflexion auf die Natur der Naturwissenschaften auch explizit umfassen." (Höttecke & Rieß, 2015, p. 130)[43]

Just as Höttecke and Rieß (2015), a number of researchers argue for the field of models and modelling that students do not understand how models are used in the development of scientific ideas because they are not enough used in this way in school (Danusso, Testa, & Vicentini, 2010; Justi & Gilbert, 2002; Khan, 2011; Krell & Krüger, 2013). Gobert et al. (2011) argue that, "as typical science instruction does not represent the real world of science and scientific practices, it is not surprising that students have naïve views of the nature of science, of scientific inquiry, and the nature of

[43] "Teachers think about the topic of experimentation from the perspective of the mediation of knowledge (Flick, 2000). [...] A clear orientation towards teaching and learning about nature of science is hardly recognizable because there is a lack of didactic strategies for dealing with openness and insecurity (Ruhrig & Höttecke, 2015). It is therefore not surprising that pupils, too, make a clear difference between experimentation in school and experimentation in science (Henke & Höttecke, 2013). If the students' own experimentation is to serve as a resource for experience in order to make students capable to reflect on experimentation in the field of research, then learning opportunities should explicitly encompass reflection on nature of science."

models (Carey et al., 1989; Driver, Leach, Millar, & Scott, 1996; Lederman, 1992)". The argument, which was derived from observations in the classroom and interrogations of teachers, is supported further by Campbell et al.'s (2015) review of research articles concerning modelling pedagogies. They found the majority of articles to contain approaches which aim at fostering conceptual understanding and thus using models for 'learning science' rather than for 'learning about science' (Hodson, 2014). If models are being used as demonstrative tools, than it is only logic that students think models should resemble the original as closely or explain how it works as exactly as possible. Reasons for why promotions concerning 'learning about models' seem to be so scarce are given by Crawford and Cullin (2004):

> „However, when pressed to elaborate on their intentions and to give examples, they [the teachers] cited time, curriculum, and technological constraints as obstacles to engaging their own students in modelling activities. Prospective teachers indicated that such time-consuming activities might interfere with what they perceived as real content." (p. 1397)

The way models are used in school is being linked by some researchers to teachers' having a rather limited meta-modelling knowledge themselves (Borrmann et al., 2014; Crawford & Cullin, 2004, 2005; Justi & van Driel, 2005; van Driel & Verloop, 1999, 2002). Windschitl and Thompson (2006) argue:

> "If teachers believe a model is an unproblematic representation of a real-world structure or process, they are less likely to value its development by students or value helping students understand the nature and function of models." (pp. 818–819)

Teachers ought not be blamed for a lack of meta-modelling knowledge, because even science textbooks fail to inform that models are human constructs that are tentative in nature (Gericke, Hagberg, & Jorde, 2013;

Harrison & Treagust, 2000; Ubben, Nitz, Rousseau, & Upmeier zu Belzen, 2015). Even if teachers were aware that models are not "unproblematic representations" of reality, they might rely on distinguishing analogies between a model and its relating original as clearly as possible (Hardwicke, 1995, p. 64) and thereby focus primarily on the model object rather than on the model itself (Mahr, 2008). Passmore et al. (2014) argue this point by saying:

> "Too great a focus on the material form of a model can be problematic because it tends to collapse the triadic relationship (between model, cognitive agent, and phenomenon) back into a dyadic one (model and phenomenon only)." (p. 1180)

6.2. Promotional Activities

The repeated observation that students' meta-modelling knowledge did increase across grades, even though only little, shows two things: (1) students' meta-modelling knowledge seems not to be a stable construct but learnable and, (2) in school, there seems to be only little or ineffective promotion of students' meta-modelling knowledge with regard to models as instruments of science. The aim of this research project was to contribute to a promotion of students' model competence by describing a baseline for interventions with the aid of an efficient diagnostic instrument for meta-modelling knowledge. In the light of this aim, the NRC (2001) invokes that "educational assessment does not exist in isolation, but must be aligned with curriculum and instruction if it is to support learning" (p. 3). In the following, I will outline propositions for a potentially successful promotion of model competence based on the findings of the present research project. As the primary focus of this research project was to develop a diagnostic instrument, the compilation makes no claim to be complete but captures elements that arose during this work.

Students' need opportunities in school to broaden their meta-modelling knowledge from a perspective of representation to a perspective of instrumentation (cf. Table 2)

With the help of the diagnostic instrument, it was possible to outline that students' meta-modelling knowledge of the nature and the purpose of models is rather strongly focussed on the perspective of representation (Gogolin & Krüger, submitted; Krell et al., 2016; Mahr, 2008). In order to broaden students' perspectives towards a perspective of instrumentation, there need to be more activities in the biology classroom concerning

‘learning about science’ (Hodson, 2014) than concerning ‘learning science’ (Hodson, 2014). Krüger, Upmeier zu Belzen, and Krell (2016) point out that an understanding about how models help scientists in their research is not being reached by proposals about how models could promote the acquisition of science content knowledge (Högermann & Kricke, 2012).
Fleige, Seegers, Upmeier zu Belzen, and Krüger (2016) as well as Günther et al. (2016) recommend for the promotion of model competence that an imperative reflection needs to stress the level III perspectives from the ‘model of model competence’. A possible realisation of supportive material for such an intervention can be found in Article 5:

Mathesius, S. & Gogolin, S. (2017). Die Letzten werden die Ersten sein. Praktisches Modellieren von Planktonkörperformen. *Biologie 5-10, 17,* 10-13.

Article 5 lays out a lesson plan for the promotion of model competence on the issue of planktons’ adaptations which hinder sinking below the oceans’ photic zone. This approach is contrary to many lesson plans that focus on which body shape will be the fastest in a swimming race (Barfod-Werner, 1992). We believed it to be likely that students are aware of the streamlined shape to enhance swimming speed and modelling in this context would rather be a medial representation of knowledge rather than an investigation based on genuine hypotheses. Therefore, we flipped the aim of the race and as a problem-orientation used the dependence of phytoplankton from light. In the promotional activity, students work in teams of two to construct plankton models from modelling clay. In this process the students hypothesise and discuss about shapes that slow sinking (possible focus in the lesson on the aspects ‘nature of models’ and ‘alternative models’). Their models then represent their hypotheses. In a second step, the students of one team race their model against those of other teams, hypothesising about the

winner (slowest wins) and changing their model based on their observations (possible focus in the lesson on the aspects ‘purpose of models’, ‘testing models’ and ‘changing models’). The reflection about the process of modelling (Günther et al., 2016) is structured and supported by a scheme for the use of models for scientific inquiry (Fleige et al., 2016). With the help of this scheme, the students visualise and conceptualise the previously pursued process of modelling for the example of plankton shapes and formulate mnemonic sentences for modelling as a scientific practice.

Additional activities to be used by teachers in their biology classrooms with an explicit focus on broadening students’ perspective of models as instruments have been laid out in Article 6:

Gogolin, S. & Krüger, D. (in press). Modellverstehen im Biologieunterricht diagnostizieren und fördern. ***Der mathematische und naturwissenschaftliche Unterricht.***

Article 6 describes five different stages (Table 12) during an intervention in an exhibition (modellSCHAU; Grotz, 2015) in the Botanical Garden Berlin.

Table 12. Stages and promoted aspects during the intervention (study VIII).

Stages (Model contexts)	Promoted Aspects	Additional suggestions for interventions in school
Reconstruction of a *T. rex*	N, T, C	– Ross, Duggan-Haas, and Allmon (2013) – Scheersoi and Dierkes (2012) – Christian (2012)
Discovery of DNA structure	N, P, T, C	– Barke and Harsch (2001) – Krell and Krüger (2016) – Zabel (2001)
Historical bio membrane models	N, A, C	– Krell, Hanauer, and Fleige (2016) – Jahnke, Austenfeld, and Lumer (2013) – Krautwig (2013)
Blackbox	N, A, P, T, C	– Krell and Reinisch (2013) – Hergert, Krell, and Krüger (in prep)

Arabidopsis thaliana as a model organism	P, T	– Gogolin and Mathesius (2014) – Mathesius and Gogolin (in press) – Ruppert (2011)

Note: N = ‘Nature of models’; A = ‘Alternative models’; P = ‘Purpose of models’; T = ‘Testing models’; C = ‘Changing models’.

All of the interventions from Table 12 can be extended and used by teachers in school, ideally in a differentiated fashion. The utilisation of the diagnostic instrument can designate students to a specific activity based on their diagnostic result and thereby help internal differentiation (Hartig et al., 2008; Ingenkamp & Lissmann, 2008). All of the activities focused on promoting students’ meta-modelling knowledge in direction of level III and thus on the use of models in a science context. On each stage, the students’ were asked to express their own perspectives about the respective model. This was done as a spontaneous alternative to the online diagnostic instrument. Approaches from research into students’ conceptions (Chi, 1992; Duit, 2002; Gropengießer, 2007; Hammann & Asshoff, 2014; Kattmann, 2015; Strike & Posner, 1992) helped subsequently to work with students’ perspectives. In case of level I perspectives, if possible, the answers were contrasted with contradicting evidence (Strike & Posner, 1992). For example, in the case of the DNA structural model, the students in the intervention at first named purposes like “visualisation”, “explanation” and “enlargement”. After watching a film about Watsons and Cricks endeavour to search for the DNA structure by means of the model, students named purposes like “research”, “discovery” and “presumptions”. On the other side, when students were asked which information about dinosaurs is certain and which is not, discussions among students almost automatically turned in direction of level III. Perspectives had only to be clarified and reinforced (Duit, 2002; Kattmann, 2015). To begin with, nearly all students believed that some characteristics of dinosaurs can be known (e.g., skeleton structure, size) and others are missing (e.g., skin colour, diet) but then some

students doubt the certainty of the knowledge about the familiar parts and tried to persuade the others. The uncertainty was enhanced by two reconstructions of T. rex, one walking upright and one walking parallel to the ground (Ross et al., 2013). In order to take account of the suggestion by Chi (1992) who stresses that students should learn the meaning of individual concepts in the context of the properties of the target ontological category, meaning in this case that students need to reflect individual concepts within authentic modelling contexts, we integrated an activity with a blackbox. The constructive approach has been used successfully to foster students' meta-modelling knowledge in both aspects 'nature of models' and 'purpose of models' by Hergert et al. (in prep) who make students' modelling activity visible with the help of a water black box which produces a specific data pattern. The students are engaged in an active cyclical modelling process where they create and change models based on the data emitted by the black box before finally reflecting about their activities and their models. They have to accept that, just like in authentic science research, there is no final clarification of the contents of the blackbox (Hergert & Krüger, 2017). Finally, teachers need to accept that any concept (e.g., meta-modelling knowledge in the aspects 'nature of models' and 'purpose of models') gets assigned to different categories (e.g., every-day life and science), where one does not replace the other but the two meanings are accessed under different circumstances (Chi, 1992).

Modelling in remote contexts enhances the promotion of meta-modelling knowledge

The promotion of students' meta-modelling knowledge might be more enhanced by some contexts than others. The qualitative findings during this research project showed that students frequently referred to level III in the aspect 'nature of models' when explaining their choice in the contexts of the

T. rex, the Neanderthal man and the evolution. The students argued that these models are theoretical reconstructions because there is uncertainty about the knowledge concerning these phenomena in general. This probably results from the fact that all these models are reconstructions of phenomena that happened partly or completely in a past time which humans have no access to. Such models may be used to introduce the perspective of the nature of models on level III. Still, teachers need to pay attention to contrasting the inability to access knowledge (because time travel has not yet been invented) with the inherent tentativeness of scientific knowledge. It needs to be reflected that the research process has no end and that there is no final state of the knowledge (Lederman & Lederman, 2014). These recommendations are in line with literature based suggestions concerning the use of historical stories about model building processes (e.g., DNA, bio membrane) in order to reflect that (1) models are scientifically accepted interpretations of a phenomenon which are plausible only in the context of the current information (Grünkorn & Fleige, 2016; Justi & Gilbert, 2002; McComas, 2008) and that (2) modelling is a fundamental knowledge construction practice in science (Bailer-Jones, 2002; Giere, 2001; Odenbaugh, 2005). Justi (2000) proposes a reflection about how and why historical models changed over time to help students get insight into epistemological processes in science. Nevertheless, it is important to address the conclusions which students draw from these historical processes and to reflect current models in order to establish science as an ongoing endeavour and models as one of its tools. Certainly, the results of study VII with regard to students' justifications of perspectives toward different model contexts can serve as a basis to develop interventions to improve teaching and learning about the nature and purpose of models. A reflection about models based on different contexts can be helpful because students' understanding of one model may be used to broaden their understanding of other models (Clough & Driver, 1986).

7. Prospective for Future Research

Science education standard documents (e.g., Germany: KMK, 2005; USA: NGSS Lead States, 2013) point out that understanding models as a form of scientific inquiry is an essential learning goal in the process of developing scientific literacy. Teachers are consequently asked to foster their students with regard to this domain. The research in the field of models and modelling in science has shown that the way from such demands to their actual realisation in school is long (Krell et al., 2016; Oh & Oh, 2011).

This research project took up the work done in a series of previous studies, starting from the setup of the 'model of model competence' (Upmeier zu Belzen & Krüger, 2010), its empirical evaluation and validation (Grünkorn, 2014; Krell, 2013; Patzke et al., 2015; Terzer, 2012; Trier et al., 2014) and its application for assessment and interventions (Fleige et al., 2016; Günther et al., 2016; Mathesius, Hartmann, Upmeier zu Belzen, & Krüger, 2016; Orsenne, 2015).

As asked for by Grünkorn (2014) in the final chapter of her dissertation, we developed an efficient and computerised diagnostic instrument for students' meta-modelling knowledge. Our forced choice tasks provide a possibility for biology teachers to detect the status quo of their students' meta-modelling knowledge in the aspects 'nature of models' and 'purpose of models' and consequently foster them individually. In the future, we can use this basis to work on issues belonging to three broad categories (1) 'improvement of the diagnostic instrument', (2) 'research into students' model competence' and (3) 'research into relations of students' meta-modelling knowledge with other constructs'.

How could the diagnostic instrument be improved to help teachers and students?

At this point in time, the diagnostic instrument comprises merely two of five aspects defined by the 'model of model competence' (Krell et al., 2016). The aspects 'alternative models', 'testing models' and 'changing models' might also be operationalised with forced choice tasks and integrated into the online version of the instrument. In order to keep the length of the diagnosis to a minimum to avoid concentrations loss (Optimizing-Satisficing-Problem; Krosnick, 1999; Moosbrugger & Kelava, 2012), teachers could be given the chance to choose which aspect to diagnose depending on the promotional activity they have planned (Günther et al., 2016). Additionally to the extension of aspects, once the influence of context-specificity of students' model competence had been investigated further, it would be sensible to develop more tasks for different model contexts. The variety of contexts would improve the coverage of the content domain and increase validity (AERA et al., 2014). In order to make the instrument more sensitive to students' individual perspectives and to increase the accuracy of the diagnosis, we could use computerised adaptive testing by enlarging the pool of tasks and administering to a student a particular sequence of tasks, based on a his or her response to a previous task (Cohen, Manion, & Morrison, 2013; Moosbrugger & Kelava, 2012).

The proposed score interpretation for the diagnostic instrument is so far only validated for students in grades ten to twelve. Whether the diagnosis produces valid results also for university students, could be easily checked by means of an additional empirical study. A pilot study which was performed with preservice science teachers ($N = 27$), who equally completed our diagnostic instrument and the Ko-WADiS questionnaire (Mathesius et al., 2016), showed that no ceiling effect is to be expected.

A multitrait-multimethod study with the Ko-WADiS questionnaire and first year university students or an instrument similar to the Ko-WADiS questionnaire but for the intended subgroup of students in school could furthermore produce valuable evidence for discriminant validity; a source that was so far not used in the present research project. One possible instrument is a questionnaire by Arnold, Kremer, and Mayer (2014) designed for students in grades eleven and twelve which includes tasks for designing an experiment. Theoretically, the questionnaire is similar to the Ko-WADiS aspect 'conducting investigations' and should thus produce discriminant validity evidence.

Turning towards the other end of the spectrum, it was possible, on the basis of the forced choice tasks, to develop tasks for students from lower grades. First steps in this direction have been taken in a collaboration project for students from primary school (grades 3-6). The task construction process and results of an empirical evaluation of the tasks have been published in Article 7:

Gogolin, S., Krell, M., Lange-Schubert, K., Hartinger, A., Upmeier zu Belzen, A., & Krüger, D. (2017). Erfassung von Modellkompetenz bei Grundschüler_innen. In H. Giest, A. Hartinger, & S. Tänzer (Eds.), *Vielperspektivität im Sachunterricht* (pp. 108–115). Bad Heilbrunn: Klinkhardt-Verlag.

In a second round, we were successful in constructing two tasks for the aspect 'nature of models' and 'purpose of models' respectively for the contexts 'dinosaur' and 'water cycle' which were understood sufficiently well by students (Lange-Schubert, Gogolin, & Krell, 2016). At the moment, this pool of tasks is being extended and will be used in combination with other instruments (modelling practices; Vo et al., 2015) in an intervention

study with primary school students by Florian Böschl (Leipzig) as part of his PhD.

What are the research gaps in the field of students' model competence that could be addressed by means of the diagnostic instrument?

The instrument may be used to systematically undertake research on students' model competence. Some further research questions may include aspects that have been targeted yet but have not been sufficiently investigated.

a. Is students' model competence global or aspect-dependent (Krell et al., 2014c)[44]?
b. How can students' meta-modelling knowledge be differentiated theoretically with regard to context-specificity (Gogolin & Krüger, 2016a; Krell et al., 2014a, 2012)?
c. Of what nature is the relationship between students' meta-modelling in the aspects 'nature of models' and 'purpose of models'? (Grünkorn, 2014; Krell, 2013; Terzer, 2012)
d. Of what nature is the relationship between students' meta-modelling knowledge and their modelling practice (Louca & Zacharia, 2012)?
e. To what extent does background information in the task stem influence students' scores of meta-modelling knowledge?
f. To what extent do difficulty generating task characteristics in the wording of the tasks influence students' scores of meta-modelling knowledge? (Kauertz, 2008; Krell, 2016)
g. How successful are the suggestions in chapter 6.2 ('Promotional Activities') to foster students' meta-modelling knowledge?

[44] The citation refers to researchers who have already worked on this issue but have reported that more research is needed.

h. Can students improve their meta-modelling knowledge without any improvement in their content knowledge (Bamberger & Davis, 2013; Czeskleba, 2016)?

Because I agree that the differentiation of the theoretical basis from the 'model of model competence' with regard to contexts is of utmost importance for assessment and teaching (Krell et al., 2014a), I would like to present a suggestion for an investigative approach for this matter.

The claim that students' meta-modelling knowledge is context-dependent has been raised repeatedly on the basis of empirical findings (Al-Balushi, 2011; Guerra-Ramos, 2012; Krell et al., 2014a, 2012; Leach, Millar, Ryder, & Séré, 2000; Lee et al., 2015; Pluta et al., 2011; Sins et al., 2009). Still, probably due to the abundance of theoretical frameworks and possible model contexts, no strictly structured study has been performed to investigate which kinds of contexts lead to which kinds of perspectives. First, the types of contexts need to be defined. In reference to Giere (2001), I propose to use four types of model contexts that proved fruitful in allowing students' perspectives on all levels: (1) original is too large (water cycle), (2) too small (virus), (3) too complex (brain) or (4) too remote (dinosaur). Second, I propose to use the category system of Grünkorn (2014) and apply it to students answers to contextualised open-ended tasks which in contrast to existing tasks (Grünkorn, 2014) do not contain a picture of the model object but merely a picture related to the original. This approach would exclude influences of the representational form of the model object (Al-Balushi, 2011; Lee et al., 2015) and might allow to detect clusters of perspectives for certain types of models.

Another approach to investigating students' response processes with regard to meta-modelling knowledge is the eye tracking method. Eye tracking is often mentioned as an alternative to think aloud protocols when gathering information about students' response processes (Gorin, 2007; Leighton,

2004; Snow & Lohman, 1989). An eye tracker can record students' eye movement during a task response. The assumption behind the approach is that the fixation of certain areas corresponds to visual attention and with that to cognitive processing resources (Gorin, 2007; Rayner, Warren, Juhasz, & Liversedge, 2004).[45] While being frequently used in cognitive psychological research, eye tracking studies which help test development in education are rather scarce (Dannemann, 2015; Gorin, 2007; Leighton, 2004; Ubben et al., 2015). A small eye tracking study that was performed with the final diagnostic instrument and students from grades eleven and twelve ($N = 20$)[46] revealed noteworthy information concerning the wording of answer options in the forced choice tasks.[47] Similar to Dannemann (2015), we used eye trackers in combination with cued retrospective reporting (Holmqvist et al., 2015). We computed heat maps which encode the fixation time of a student on a certain area into a colour scheme (SMI SensoMotoric Instruments, 2016). This way, it is possible to detect signal words that triggered the students' attention (fixation time > 300 ms) and thus may act as formal difficulty generating task characteristics (Dannemann, 2015; cf. Kauertz, 2008). The findings were rather uneventful except for the context of the evolution in the aspect 'purpose of models'. There, the word 'vereinfacht' in the answer option on level I which specifies the operator 'abbilden' causes a long fixation time (> 300 ms;
Figure 5) and can be seen as a formal difficulty generating task characteristic as this adverb has no significance for the aspect 'purpose of models' but seemingly leads students to choose the level I answer option.

[45] It has to be remarked that, although some researchers produced empirical connections between visual fixations and cognitive processing (Holmqvist, Nyström, & Andersson, 2015; Rayner et al., 2004), "direct evidence that the location of an individuals' gaze corresponds directly to the information being processed by the individual is still needed" (Gorin 2006, p. 29).

[46] Christoph van Heteren-Frese reports some findings of the eye tracking study in his master thesis (van Heteren-Frese, 2016).

[47] Thanks to Sascha Tamm from the department of experimental and neurocognitive psychology for his time and his support during the study.

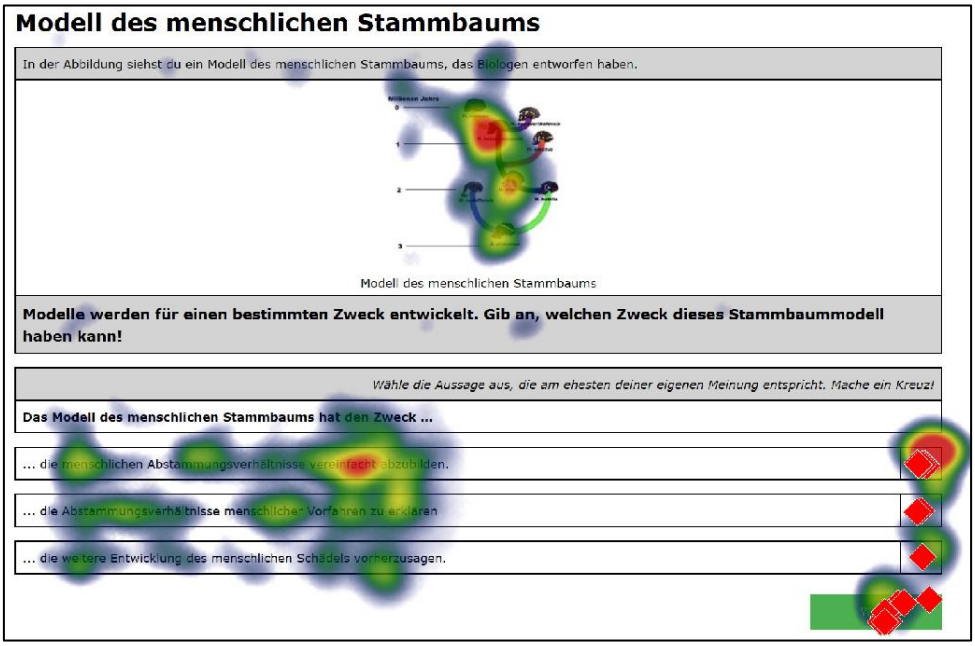


Figure 5. Heat map for the context evolution in the aspect 'purpose of models'. *N* = 10, Scale: 50 ms (blue) – 368 ms (red). The red squares indicate the clicks of the students.

The data from the verbal reports underlines this observation: "deswegen habe ich das einfach als Antwort gewählt, dass das eine Vereinfachung ist."[48]. The adverb 'vereinfachen' rather denotes a perspective within the aspect 'nature of models' where on level II models may be seen as idealised representations. This poses a problem for a valid score interpretation as the students' perspectives may in this case not be diagnosed in a clear distinction. Ideally, the adverb should be removed from the answer option and all tasks should be tested again.

The method of eye tracking turned out to be a fruitful approach to integrating students' response processes into score validation for our forced choice tasks. Future research using this method may produce valuable insights into students' response processes for the purposes of both learning

[48] "that's why I chose as an answer that it [the model] is a simplification.". [Exerpt from a report by one of the students who fixated the adverb 'vereinfacht' for more than 300 ms.]

about their meta-modelling knowledge and gathering evidence for valid score interpretations.

How can the diagnostic instrument be used to detect relationships of students' model competence with other constructs?

Contemporary research in science education has a strong focus on scientific inquiry and nature of science (Lederman & Abell, 2014). As methods of scientific inquiry, Mayer (2007) mentions the following: observing, investigating, describing, comparing, classifying, experimenting and using models. Studies have shown that, no matter what domain of scientific inquiry, students have difficulties understanding how to solve biological problems successfully (experimenting: Hammann, Phan, Ehmer, & Grimm, 2008; comparing: Krüger & Burmester, 2005; observing: Wellnitz, 2012; using models: Krell et al., 2016).

Considering there to be "a general pattern to all scientific reasoning" (Giere, Bickle, & Mauldin, 2006, p. 6), it ought to be of interest for science education researchers of what nature the relationships between the different methods of inquiry are. An approach to assessing scientific reasoning skills across different methods of inquiry ('conducting investigations' and 'using models') and different subjects (Biology, Chemistry, Physics) has been provided by Mathesius et al. (2016) for preservice science teachers. A related approach for school students is provided by the VerE-Study (Nowak, Nehring, Tiemann, & Upmeier zu Belzen, 2013). The results of this approach are reportedly valid only for students of grades nine and ten (Nowak et al., 2013). Combining the diagnostic instrument from the present research project with the instrument of Arnold, Kremer, and Mayer (2010) would allow to investigate students from grades ten to twelve with reference to both methods of inquiry. The combination with the other just mentioned instruments would eventually grant more insight into the process of

development of the patterns of scientific reasoning (Giere et al., 2006) across years in school and in university.

The instrument of Mathesius et al. (2016) could also be combined with the diagnostic instrument of the present research project to relate teachers' scientific reasoning in the domain of 'using models' with their students' meta-modelling knowledge in order to investigate the frequently raised hypothesis that the former determines the latter (Grünkorn, 2014; Krell, 2013; Terzer, 2012).

Last but not least, the question of whether an elaborate understanding of nature of science aspects is a precondition for understanding the nature of models to be tentative and the role of models as instruments of scientists (Chi, 1992) or, reverse, if an elaborate understanding of the nature of models reinforces the understanding of nature of science aspects (Gobert et al., 2011) could be investigated by combining the diagnostic instrument with another instrument intended to measure nature of science aspects, as, for example, the 'Views of nature of science questionnaire' by Lederman, Abd-El-Khalick, Bell, and Schwartz (2002) and interventions of model competence (Fleige et al., 2016) as well as interventions of nature of science aspects (Abd-El-Khalick, 2013).

8. References

Abd-El-Khalick, F. (2013). Teaching With and About Nature of Science, and Science Teacher Knowledge Domains. *Science & Education*, *22*(9), 2087–2107. https://doi.org/10.1007/s11191-012-9520-2

Acher, A., Arcà, M., & Sanmartí, N. (2007). Modeling as a teaching learning process for understanding materials: A case study in primary education. *Science Education*, *91*(3), 398–418. https://doi.org/10.1002/sce.20196

Adams, R. J. (2005). Reliability as a measurement design effect. *Studies in Educational Evaluation*, *31*(2-3), 162–172. https://doi.org/10.1016/j.stueduc.2005.05.008

Adams, W. K., & Wieman, C. E. (2011). Development and Validation of Instruments to Measure Learning of Expert-Like Thinking. *International Journal of Science Education*, *33*(9), 1289–1312. https://doi.org/10.1080/09500693.2010.512369

Adúriz-Bravo, A. (2013). A 'Semantic' View of Scientific Models for Science Education. *Science & Education*, *22*(7), 1593–1611.

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

Aikenhead, G. S., & Ryan, A. G. (1992). The Development of a New Instrument: 'Views on Science-Technology-Society' (VOSTS). *Science Education*, *76*(5), 477–491.

Ainsworth, S. (2008). The educational value of multiple-representations when learning complex scientific concepts. In J. K. Gilbert (Ed.), *Models*

and modeling in science education. Visualization in science education (pp. 191–208). Dordrecht: Springer.

Al-Balushi, S. (2011). Students' evaluation of the credibility of scientific models that represent natural entities and phenomena. *International Journal of Science and Mathematics Education, 9*(3), 571–601.

Anastasi, A. (1976). *Psychological testing*. New York: Macmillan.

Arnold, J. C., Kremer, K., & Mayer, J. (2014). Understanding Students' Experiments - What kind of support do they need in inquiry tasks? *International Journal of Science Education, 36*(16), 2719–2749. https://doi.org/10.1080/09500693.2014.930209

Arnold, J., Kremer, K., & Mayer, J. (2010). Wissenschaftliches Denken beim Experimentieren – Kompetenzdiagnose in der Sekundarstufe II. *Erkenntnisweg Biologiedidaktik, 11*, 7–20.

Aufschnaiter, C., Cappell, J., Dübbelde, G., Mayer, J., Möller, A., Stiensmeier-Pelster, J., & Wolgast, A. (2012). Assessing Prospective Teachers' Diagnostic Competence. In C. Bruguière, A. Tiberghien, & P. Clément (Eds.), *E-Book Proceedings of the ESERA 2011 Conference: Science learning and Citizenship.* (pp. 125–131). Lyon: European Science Education Research Association.

Bailer-Jones, D. (1999). Tracing the Development of Models in the Philosophy of Science. In L. Magnani, N. J. Nersessian, & P. Thagard (Eds.), *Model-Based Reasoning in Scientific Discovery* (pp. 23–40). Boston: Springer. https://doi.org/10.1007/978-1-4615-4813-3_2

Bailer-Jones, D. (2002). Scientists' thoughts on scientific models. *Perspectives on Science, 10*(3), 275–301.

Bailer-Jones, D. (2003). When scientific models represent. *International studies in the philosophy of science, 17*(1), 59–74.

Bailer-Jones, D. (2009). *Scientific Models in Philosophy of Science*. Pittsburgh: University of Pittsburgh Press.

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*(2), 213–238. https://doi.org/10.1080/09500693.2011.624133

Barfod-Werner, I. (1992). Wettschwimmen der Wachsformen. *Unterricht Biologie*, *16*(178), 14–16.

Barke, H.-D., & Harsch, G. (2001). Watson und Crick: Nobelpreisträger spielen mit Modellen. In H.-D. Barke & G. Harsch (Eds.), *Chemiedidaktik Heute. Lernprozesse in Theorie und Praxis* (pp. 485–496). Berlin: Springer.

Barnhart, C. (Ed.). (1953). *The American College Dictionary*. New York: Harper & Brothers.

Baxter, G. P., & Glaser, R. (1998). Investigating the Cognitive Complexity of Science Assessments. *Educational Measurement: Issues and Practice*, *17*(3), 37–45. https://doi.org/10.1111/j.1745-3992.1998.tb00627.x

Bennett, R. E. (1991). On the meaning of constructed response. *ETS Research Report Series*, *2*, 1–46. https://doi.org/10.1002/j.2333-8504.1991.tb01429.x

Berland, L., & Crucet, K. (2016). Epistemological Trade-Offs: Accounting for Context When Evaluating Epistemological Sophistication of Student Engagement in Scientific Practices. *Science Education*, *100*(1), 5–29. https://doi.org/10.1002/sce.21196

Böckenholt, U. (2004). Comparative judgments as an alternative to ratings: identifying the scale origin. *Psychological Methods*, *9*(4), 453–465. https://doi.org/10.1037/1082-989X.9.4.453

Bond, T. G., & Fox, M. C. (2001). *Applying the Rasch Model*. Mahwah, N.J.: Erlbaum.

Borrmann, J., Reinhardt, N., Krell, M., & Krüger, D. (2014). Perspektiven von Lehrkräften über Modelle in den Naturwissenschaften: Eine generalisierende Replikationsstudie. *Erkenntnisweg Biologiedidaktik, 13*, 57–72.

Borsboom, D., & Markus, K. A. (2013). Truth and Evidence in Validity Theory. *Journal of educational measurement*, *50*(1), 110–114.

Boulter, C. J., & Buckley, B. C. (2000). Constructing a typology of models for science education. In J. K. Gilbert & C. J. Boulter (Eds.), *Developing Models in Science Education* (pp. 41–57). Dordrecht: Springer.

Boumans, M. (1999). Built-in justification. In L. Magnani, N. J. Nersessian, & P. Thagard (Eds.), *Model-Based Reasoning in Scientific Discovery* (pp. 66–96). Boston: Springer.

Brennan, R., & Prediger, D. (1981). Coefficient Kappa: Some uses, misuses, and alternatives. *Educational and Psychological Measurement*, *41*, 687–699.

Briggs, D., Alonzo, A., Schwab, C., & Wilson, M. (2006). Diagnostic Assessment With Ordered Multiple-Choice Items. *Educational Assessment*, *11*(1), 33–63. https://doi.org/10.1207/s15326977ea1101_2

Burnham, K. P., & Anderson, D. R. (2004). Multimodel inference understanding AIC and BIC in model selection. *Sociological Methods & Research*, *33*(2), 261–304.

Campbell, D. T., & Fiske, D. W. (1959). Convergent and discriminant validation by the multitrait-multimethod matrix. *Psychological Bulletin*, *56*(2), 81–105.

Campbell, T., & Oh, P. S. (2015). Engaging Students in Modeling as an Epistemic Practice of Science: An Introduction to the Special Issue of the

Journal of Science Education and Technology. *Journal of Science Education and Technology*, *24*(2-3), 125–131. https://doi.org/10.1007/s10956-014-9544-2

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*(1), 159–176.

Campbell, T., Oh, P., & Neilson, D. (2013). Reification of five types of modeling pedagogies with model-based inquiry (MBI) modules for high school science classrooms. In M. S. Khine & I. M. Saleh (Eds.), *Approaches and strategies in next generation science learning* (pp. 106–126). Hershey, PA: Information Science Reference.

Campbell, T., Schwarz, C., & Windschitl, M. (2016). What We Call Misconceptions May Be Necessary Stepping-Stones Toward Making Sense of the World. *Science and Children*, *53*(7), 69–74.

Carey, S., Evans, R., Honda, M., Jay, E., & Unger, C. (1989). 'An experiment is when you try it and see if it works': A study of grade 7 students' understanding of the construction of scientific knowledge. *International Journal of Science Education*, *11*(5), 514–529. https://doi.org/10.1080/0950069890110504

Cartier, J., Rudolph, J., & Stewart, J. (2001). The Nature and Structure of Scientific Models. The National Center for Improving Student Learning and Achievement. Retrieved from http://biology.westfield.ma.edu/biol104w/sites/default/files/Models.pdf

Chaoui, N. (2011). *Finding Relationships Between Multiple-Choice Math Tests And Their Stem-Equivalent Constructed Responses*. Claremont Graduate University: ProQuest Dissertations Publishing.

Chi, M. (1992). Conceptual Change within and across Ontological Categories: Examples from Learning and Discovery in Science. In R.

Giere & H. Feigl (Eds.), *Cognitive Models of Science* (pp. 129–186). Minneapolis: University of Minnesota Press.

Chittleborough, G. D., Treagust, D. D., 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 and Technological Education, 23*, 195–212.

Choy, L. T. (2014). The strengths and weaknesses of research methodology: comparison and complimentary between qualitative and quantitative approaches. *IOSR Journal of Humanities and Social Science*, *19*(4), 99–104.

Christian, A. (2012). Schlappe Schleicher oder rasante Renner? *Unterricht Biologie*, *374*, 42–48.

Clement, J. (1989). Learning via Model Construction and Criticism. In J. A. Glover, R. R. Ronning, & C. R. Reynolds (Eds.), *Perspectives on Individual Differences. Handbook of Creativity.* Boston: Springer.

Clement, J. J., & Rea-Ramirez, M. A. (Eds.). (2008). *Models and modeling in science education: Vol. 2. Model Based Learning and Instruction in Science*. Dordrecht: Springer.

Clough, E. E., & Driver, R. (1986). A study of consistency in the use of students' conceptual frameworks across different task contexts. *Science Education, 70*(4), 473–496.

Cohen, L., Manion, L., & Morrison, K. (2013). *Research methods in education* (7th ed.). New York: Routledge.

Cohors-Fresenborg, E., Sjuts, J., & Sommer, N. (2004). Komplexität von Denkvorgängen und Formalisierung von Wissen. In M. Neubrand (Ed.), *Mathematische Kompetenzen von Schülerinnen und Schülern in Deutschland. Vertiefende Analysen im Rahmen von PISA 2000* (pp. 109–144). Wiesbaden: VS Verlag für Sozialwissenschaften.

Coll, R. K. (2006). The Role of Models, Mental Models and Analogies in Chemistry Teaching. In P. J. Aubusson, A. G. Harrison, & S. M. Ritchie (Eds.), *Science & Technology Education Library: Vol. 30. Metaphor and Analogy in Science Education* (pp. 65–77). Dordrecht: Springer.

Cooper, L. G. (1983). A review of multidimensional scaling in marketing research. *Applied Psychological Measurement*, *7*(4), 427–450.

Cortina, J. M. (1993). What is coefficient alpha? An examination of theory and applications. *Journal of applied psychology*, *78*(1), 98–104.

Crawford, B. A., & Cullin, M. J. (2004). Supporting prospective teachers' conceptions of modelling in science. *International Journal of Science Education*, *26*(11), 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. Eijkelhoff (Eds.), *Research and the quality of science education* (pp. 309–323). Dordrecht: Springer.

Creswell, J., & Plano Clark, V. (2011). *Designing and conducting mixed methods research*. Los Angeles, CA: Sage.

Crocker, L., & Algina, J. (1986). *Introduction to classical and modern test theory*. Orlando: ERIC.

Cronbach, L. J. (1951). Coefficient alpha and the internal structure of tests. *Psychometrika*, *16*(3), 297–334.

Cronbach, L. J., & Meehl, P. E. (1955). Construct validity in psychological tests. *Psychological Bulletin, 52*(4), 281–302.

Cronbach, L. J., & Shavelson, R. J. (2004). My Current Thoughts on Coefficient Alpha and Successor Procedures. *Educational and Psychological Measurement, 64*(3), 391–418. https://doi.org/10.1177/0013164404266386

Czeskleba, A. (2016). *Bedeutung von biologischen Fachinformationen für das Lernen von Metamodeling Knowledge*. Berlin: Logos.

Dannemann, S. (2015). *Schülervorstellungen zur visuellen Wahrnehmung: Entwicklung und Evaluation eines Diagnoseinstruments*. *Beiträge zur didaktischen Rekonstruktion*. Baltmannsweiler: Schneider Hohengehren.

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*(7), 871–905. https://doi.org/10.1080/09500690902833221

Döring, N., & Bortz, J. (2016). *Forschungsmethoden und Evaluation in den Sozial- und Humanwissenschaften* (5th ed.). Berlin: Springer.

Driver, R., Leach, J., Millar, R., & Scott, P. (1996). *Young people's images of science.* Buckingham: Open University Press.

Duit, R. (2002). Alltagsvorstellungen und Physik lernen. In E. Kircher & W. Schneider (Eds.), *Physikdidaktik in der Praxis* (pp. 1–26). Berlin: Springer.

Duit, R., Gropengießer, H., Kattmann, U., Komorek, M., & Parchmann, I. (2012). The Model Of Educational Reconstruction – A Framework For Improving Teaching And Learning Science. In D. Jorde & J. Dillon (Eds.), *Science Education Research and Practice in Europe* (pp. 13–37). Rotterdam: SensePublishers.

Elby, A., & Hammer, D. (2001). On the substance of a sophisticated epistemology. *Science Education, 85*(5), 554–567.

Embretson, S. E., & Reise, S. P. (2000). *Item Response Theory for Psychologists*. Mahwah, NJ: Lawrence Erlbaum.

Ericsson, K. A., & Simon, H. A. (1980). Verbal reports as data. *Psychological Review, 87*(3), 215–251.

Ericsson, K. A., & Simon, H. A. (1993). *Protocol analysis: Verbal Reports as Data*. Cambridge, MA: MIT Press.

Ericsson, K. A., & Simon, H. A. (1998). How to study thinking in everyday life: Contrasting think-aloud protocols with descriptions and explanations of thinking. *Mind, Culture, and Activity*, *5*(3), 178–186.

Famularo, L. (2007). *The effect of response format and test taking strategies on item difficulty: A comparison of stem-equivalent multiple-choice and constructed-response test items*. Boston: ProQuest Dissertations Publishing.

Field, A. (2013). *Discovering statistics using IBM SPSS statistics: And sex and drugs and rock 'n' roll* (4th edition). Los Angeles: Sage.

Fleige, J., Seegers, A., Upmeier zu Belzen, A., & Krüger, D. (2012). Förderung von Modellkompetenz im Biologieunterricht. *Der mathematische und naturwissenschaftliche Unterricht*, *65*(1), 19–28.

Fleige, J., Seegers, A., Upmeier zu Belzen, A., & Krüger, D. (Eds.). (2016). *Modellkompetenz im Biologieunterricht Klasse 7-10*. Donauwörth: Auer.

Fleischer, J., Koeppen, K., Kenk, M., Klieme, E., & Leutner, D. (2013). Kompetenzmodellierung: Struktur, Konzepte und Forschungszugänge des DFG-Schwerpunktprogramms. *Zeitschrift für Erziehungswissenschaft*, *16*(1), 5–22.

Flick, L. B. (2000). Cognitive scaffolding that fosters scientific inquiry in middle level science. *Journal of Science Teacher Education*, *11*(2), 109–129.

Floyd, R. G., Phaneuf, R. L., & Wilczynski, S. M. (2005). Measurement properties of indirect assessment methods for functional behavioral assessment: A review of research. *School Psychology Review*, *34*(1), 58–73.

Fonteyn, M. E., Kuipers, B., & Grobe, S. J. (1993). A Description of Think Aloud Method and Protocol Analysis. *Qualitative Health Research*, *3*(4), 430–441. https://doi.org/10.1177/104973239300300403

Frigg, R., & Hartmann, S. (2006). Scientific models. In S. Sarkar & J. Pfeifer (Eds.), *Philosophy of Science. An Encyclopedia* (Vol. 2, pp. 740–749). New York: Routledge.

Gericke, N., Hagberg, M., & Jorde, D. (2013). Upper Secondary Students' Understanding of the Use of Multiple Models in Biology Textbooks—The Importance of Conceptual Variation and Incommensurability. *Research in Science Education*, *43*(2), 755–780. https://doi.org/10.1007/s11165-012-9288-z

Gibbs, A., & Lawson, A. (1992). The Nature of Scientific Thinking as Reflected by the Work of Biologists & by Biology Textbooks. *The American Biology Teacher*, *54*, 137–152.

Giere, R. N. (2001). A New Framework for Teaching Scientific Reasoning. *Argumentation*, *15*, 21–33.

Giere, R. N. (2004). How Models Are Used to Represent Reality. *Philosophy of Science*, *71*(5), 742–752. https://doi.org/10.1086/425063

Giere, R. N., Bickle, J., & Mauldin, R. F. (2006). *Understanding scientific reasoning*. Belmont: Thomson Wadsworth.

Gilbert, J. K. (2004). Models and modelling: Routes to more authentic science education. *International Journal of Science and Mathematics Education*, *2*(2), 115–130.

Gilbert, J. K., & Justi, R. (2016). *Modelling-based teaching in science education*. *Models and modeling in science education: Vol. 9*. Switzerland: Springer.

Gilbert, J. K., Boulter, C. J., & Elmer, R. (2000). *Positioning Models in Science Education and in Design and Technology Education*. Dordrecht: Springer.

Gobert, J., O'Dwyer, L., Horwitz, P., Buckley, B., 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*(5), 653–684. https://doi.org/10.1080/09500691003720671

Gogolin, S., & Krüger, D. (2015). Nature of models - Entwicklung von Diagnoseaufgaben. In M. Hammann, J. Mayer, & N. Wellnitz (Eds.), *Lehr- und Lernforschung in der Biologiedidaktik* (6th ed., pp. 27–41). Innsbruck: Studienverlag.

Gogolin, S. & Krüger, D. (2016a). Diagnosing Students' Understanding of the Nature of Models. *Research in Science Education, 1-23,* doi 10.1007/s11165-016-9551-9.

Gogolin, S., & Krüger, D. (2016b). Konstruktion von Diagnoseaufgaben zum Zweck von Modellen. *Biologie Lehren und Lernen – Zeitschrift für Didaktik der Biologie*, *1*(20), 44–62.

Gogolin, S., & Krüger, D. (in press). Modellverstehen im Biologieunterricht diagnostizieren und fördern. *Der mathematische und naturwissenschaftliche Unterricht*.

Gogolin, S. & Krüger, D. (submitted). Students' understanding of the nature and purpose of models. *Journal of Research in Science Teaching*.

Gogolin, S., & Mathesius, S. (2014). Gleich und gleich gesellt sich gern - oder nicht? *Unterricht Biologie*, *394*, 21–25.

Gogolin, S., Krell, M., Lange-Schubert, K., Hartinger, A., Upmeier zu Belzen, A., & Krüger, D. (2017). Erfassung von Modellkompetenz bei Grundschüler_innen. In H. Giest, A. Hartinger, & S. Tänzer (Eds.),

Vielperspektivität im Sachunterricht (pp. 108–115). Bad Heilbrunn: Klinkhardt-Verlag.

Gorin, J. S. (2006). Test Design with Cognition in Mind. *Educational Measurement: Issues and Practice*, *25*(4), 21–35.

Gorin, J. S. (2007). Test Construction and Diagnostic Testing. In J. Leighton & M. Gierl (Eds.), *Cognitive Diagnostic Assessment for Education. Theory and Applications* (pp. 173–202). Cambridge: Cambridge University Press. https://doi.org/10.1017/CBO9780511611186.007

Gropengießer, H. (2007). Theorie des erfahrungsbasierten Verstehens. In D. Krüger & H. Vogt (Eds.), *Theorien in der biologiedidaktischen Forschung. Ein Handbuch für Lehramtsstudenten und Doktoranden* (pp. 105–116). Berlin: Springer.

Grosslight, L., Jay, E., Unger, C., & Smith. (1991). Understanding models and their use in science: Conceptions of middle and high school students and experts. *Journal of Research in Science Teaching*, *28*(9), 799–822. https://doi.org/10.1002/tea.3660280907

Grotz, K. (2015). *Modellschau. Perspektiven auf botanische Modelle*. Ausstellungskatalog. Berlin: Laserline.

Grünkorn, J. (2014). *Modellkompetenz im Biologieunterricht: Empirische Analyse von Modellkompetenz bei Schülerinnen und Schülern der Sekundarstufe I mit Aufgaben im offenen Antwortformat*. Dissertation. Retrieved from http://www.diss.fu-berlin.de/diss/receive/FUDISS_thesis_000000097320

Grünkorn, J., & Fleige, J. (2016). Bau und Funktion der Fischhaut. In J. Fleige, A. Seegers, A. Upmeier zu Belzen, & D. Krüger (Eds.), *Modellkompetenz im Biologieunterricht Klasse 7-10* (pp. 23–28). Donauwörth: Auer.

Grünkorn, J., Lotz, A., & Terzer, E. (2014). Erfassung von Modellkompetenz im Biologieunterricht. *Der mathematische und naturwissenschaftliche Unterricht, 67*(3), 132–138.

Grünkorn, J., Upmeier zu Belzen, A., & Krüger, D. (2014). Assessing and structuring students' perspectives on biological models and their use in science to evaluate a theoretical cognitive model. *International Journal of Science Education, 36*, 1651–1684.

Guerra-Ramos, M. T. (2012). Teachers' Ideas About the Nature of Science: A Critical Analysis of Research Approaches and Their Contribution to Pedagogical Practice. *Science & Education*, *21*(5), 631–655. https://doi.org/10.1007/s11191-011-9395-7

Gulliksen, H. (1950). Intrinsic validity. *American psychologist*, *5*(10), 511–517. https://doi.org/10.1037/h0054604

Günther, S. L., Fleige, J., Upmeier zu Belzen, A., & Krüger, D. (2016). Interventionsstudie mit angehenden Lehrkräften zur Förderung von Modellkompetenz im Unterrichtsfach Biologie. In C. Gräsel & K. Trempler (Eds.), *Entwicklung von Professionalität pädagogischen Personals. Interdisziplinäre Betrachtungen, Befunde und Perspektiven.* (pp. 215–236). Wiesbaden: Springer.

Halloun, I. A. (2007). Mediated Modeling in Science Education. *Science & Education, 16*(7-8), 653–697. https://doi.org/10.1007/s11191-006-9004-3

Hammann, M., & Asshoff, R. (2014). *Schülervorstellungen im Biologieunterricht*. Stuttgart: Klett.

Hammann, M., Phan, T. T. H., Ehmer, M., & Grimm, T. (2008). Assessing pupils' skills in experimentation. *Journal of Biological Education, 42*(2), 66–72.

Hardwicke, A. J. (1995). Using Molecular Models to Teach Chemistry. Part 2: Using Models. *School science review*, *77*(279), 47–56.

Harré, R. (1988). Where models and analogies really count. *International studies in the philosophy of science*, *2*(2), 118–133. https://doi.org/10.1080/02698598808573310

Harrison, A. G., & Treagust, D. F. (1996). Secondary students' mental models of atoms and molecules: Implications for teaching chemistry. *Science Education*, *80*(5), 509–534.

Harrison, A. G., & Treagust, D. F. (2000). A typology of school science models. *International Journal of Science Education*, *22*(9), 1011–1026.

Hartig, J., & Frey, A. (2012). Konstruktvalidierung und Skalenbeschreibung in der Kompetenzdiagnostik durch die Vorhersage von Aufgabenschwierigkeiten. *Psychologische Rundschau*, *63*(1), 43–49. https://doi.org/10.1026/0033-3042/a000109

Hartig, J., & Klieme, E. (Eds.). (2007). *Möglichkeiten und Voraussetzungen technologiebasierter Kompetenzdiagnostik: Eine Expertise im Auftrag des Bundesministeriums für Bildung und Forschung*. Bonn: BMBF.

Hartig, J., Klieme, E., & Leutner, D. (2008). *Assessment of competencies in educational contexts: State of the art and future prospects*. Göttingen: Hogrefe & Huber.

Henke, A., & Höttecke, D. (2013). Entwicklung von Schülervorstellungen zur Natur der Naturwissenschaften im Rahmen forschenden Lernens und historischer Fallstudien. In S. Bernholt (Ed.), *Zur Didaktik der Chemie und Physik. GDCP-Jahrestagung in Hannover 2012.* Universität Hannover.

Henson, J. M., Reise, S. P., & Kim, K. H. (2007). Detecting mixtures from structural model differences using latent variable mixture modeling: A comparison of relative model fit statistics. *Structural Equation Modeling: A Multidisciplinary Journal*, *14*(2), 202–226.

Henze, I., van Driel, J. H., & Verloop, N. (2008). Development of Experienced Science Teachers' Pedagogical Content Knowledge of Models of the Solar System and the Universe. *International Journal of Science Education*, *30*(10), 1321–1342. https://doi.org/10.1080/09500690802187017

Hergert, S., & Krüger, D. (2017). Die Büchse der Pandora oder warum man die Blackbox nicht öffnen sollte. *Der mathematische und naturwissenschaftliche Unterricht*, *2*, 132–133.

Hergert, S., Krell, M., & Krüger, D. (in prep). How students engage in modelling by using a black box.

Hicks, L. E. (1970). Some properties of ipsative, normative, and forced-choice normative measures. *Psychological Bulletin*, *74*(3), 167–184. https://doi.org/10.1037/h0029780

Hodson, D. (2014). Learning Science, Learning about Science, Doing Science: Different goals demand different learning methods. *International Journal of Science Education*, *36*(15), 2534–2553. https://doi.org/10.1080/09500693.2014.899722

Hofer, B. K., & Pintrich, P. R. (1997). The development of epistemological theories: Beliefs about knowledge and knowing and their relation to learning. *Review of Educational Research*, *67*(1), 88–140.

Hogan, K. (2000). Exploring a process view of students' knowledge about the nature of science. *Science Education*, *84*(1), 51–70.

Högermann, C., & Kricke, W. (2012). *Modelle für den Biologieunterricht*. Hallbergmoos: Aulis.

Holmqvist, K., Nyström, M., & Andersson, R. (2015). *Eye Tracking: A comprehensive guide to methods and measures*. Oxford: Oxford University Press.

Höttecke, D., & Rieß, F. (2015). Naturwissenschaftliches Experimentieren im Lichte der jüngeren Wissenschaftsforschung – Auf der Suche nach einem authentischen Experimentbegriff der Fachdidaktik. *Zeitschrift für Didaktik der Naturwissenschaften*, *21*(1), 127–139.

Ingenkamp, K., & Lissmann, U. (2008). *Lehrbuch der pädagogischen Diagnostik* (6th ed.). *Studium Paedagogik*. Weinheim: Beltz.

Jahnke, L., Austenfeld, U., & Lumer, J. (2013). Transportvorgänge durch Membranen. *Unterricht Biologie*, *387/388*, 53–59.

Johnson, R. B., & Onwuegbuzie, A. J. (2004). Mixed Methods Research: A Research Paradigm Whose Time Has Come. *Educational Researcher*, *33*(7), 14–26. https://doi.org/10.3102/0013189X033007014

Justi, R. (2000). Teaching with historical models. In J. K. Gilbert & C. J. Boulter (Eds.), *Developing Models in Science Education* (pp. 209–226). Dordrecht: Springer.

Justi, R. S., & Gilbert, J. K. (2002). Science teachers' knowledge about and attitudes towards the use of models and modelling in learning science. *International Journal of Science Education*, *24*(12), 1273–1292. https://doi.org/10.1080/09500690210163198

Justi, R. S., & Gilbert, J. K. (2003). Teachers' views on the nature of models. *International Journal of Science Education*, *25*(11), 1369–1386. https://doi.org/10.1080/0950069032000070324

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*(5), 549–573. https://doi.org/10.1080/0950069042000323773

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* (2nd ed., pp. 64–80). New York: Routledge.

Karabenick, S. A., Woolley, M. E., Friedel, J. M., Ammon, B. V., Blazevski, J., Bonney, C. R.,. . . Kempler, T. M. (2007). Cognitive processing of self-report items in educational research: Do they think what we mean? *Educational Psychologist*, *42*(3), 139–151.

Kattmann, U. (2015). *Schüler besser verstehen: Alltagsvorstellungen im Biologieunterricht*. München: Aulis.

Kauertz, A. (2008). *Schwierigkeitserzeugende Merkmale physikalischer Leistungsaufgaben*. Berlin: Logos.

Kauertz, A., Löffler, P., & Fischer, H. E. (2010). Physikaufgaben. In E. Kircher, R. Girwidz, & P. Häußler (Eds.), *Physikdidaktik. Theorie und Praxis.* Berlin: Springer.

Kauertz, A., Neumann, K., & Haertig, H. (2012). Competence in Science Education. In B. J. Fraser, K. Tobin, & C. J. McRobbie (Eds.), *Springer International Handbooks of Education: Vol. 24. Second International Handbook of Science Education* (pp. 711–721). Dordrecht: Springer.

Kawasaki, K., Rupert-Herrenkohl, L., & Yeary, S. A. (2004). Theory building and modeling in a sinking and floating unit: A case study of third and fourth grade students' developing epistemologies of science. *International Journal of Science Education*, *26*(11), 1299–1324.

Khan, S. (2011). What's Missing in Model-Based Teaching. *Journal of Science Teacher Education*, *22*(6), 535–560.

Klieme, E., & Hartig, J. (2007). Kompetenzkonzepte in den Sozialwissenschaften und im erziehungswissenschaftlichen Diskurs. *Zeitschrift für Erziehungswissenschaft*, *10*(8), 11–29.

KMK. (2005). Standards für die Lehrerbildung: Bildungswissenschaften.: Beschluss der Kultusministerkonferenz vom 16.12.2004. *Zeitschrift für*

Pädagogik, *51*(2), 280–290. Retrieved from http://www.pedocs.de/volltexte/2011/4756/

Knoblich, G., & Öllinger, M. (2006). Die Methode des Lauten Denkens. In J. Funke & P. Frensch (Eds.), *Handbuch Allgemeine Psychologie–Kognition* (pp. 691–696). Göttingen: Hogrefe.

Koeppen, K., Hartig, J., Klieme, E., & Leutner, D. (2008). Current Issues in Competence Modeling and Assessment. *Zeitschrift für Psychologie*, *216*(12), 61–73.

Konrad, K. (2010). Lautes Denken. In G. Mey & K. Mruck (Eds.), *Handbuch Qualitative Forschung in der Psychologie*. Wiesbaden: VS Verlag für Sozialwissenschaften.

Krautwig, D. (2013). Myelinfiguren. *Unterricht Biologie*, *390*, 48–49.

Krell, M. (2013). *Wie Schülerinnen und Schüler biologische Modelle verstehen*. Berlin: Logos.

Krell, M. (2016). *What makes biological experiments difficult for students?*, Paper presented at the Conference of European Researchers in Didactics of Biology, Karlstad, Sweden.

Krell, M., & Krüger, D. (2011). Forced Choice-Aufgaben zur Evaluation von Modellkompetenz im Biologieunterricht: Empirische Überprüfung konstrukt- und merkmalsbezogener Teilkompetenzen. *Erkenntnisweg Biologiedidaktik*, *10*, 53–68.

Krell, M., & Krüger, D. (2013). Wie werden Modelle im Biologieunterricht eingesetzt? Ergebnisse einer Fragebogenstudie. *Erkenntnisweg Biologiedidaktik, 12*, 9–26.

Krell, M., & Krüger, D. (2016). Entdeckung der DNS-Struktur. In J. Fleige, A. Seegers, A. Upmeier zu Belzen, & D. Krüger (Eds.), *Modellkompetenz im Biologieunterricht Klasse 7-10* (pp. 49–57). Donauwörth: Auer.

Krell, M., & Reinisch, B. (2013). Rätsel um die schwarze Kiste: Mit der Blackbox naturwissenschaftliche Modellbildung verstehen. *Grundschule*, *45*, 16–17.

Krell, M., Czeskleba, A., & Krüger, D. (2012). Validierung von Forced Choice-Aufgaben durch Lautes Denken. *Erkenntnisweg Biologiedidaktik*, *11*, 53–70.

Krell, M., Hanauer, N., & Fleige, J. (2016). Biomembran. In J. Fleige, A. Seegers, A. Upmeier zu Belzen, & D. Krüger (Eds.), *Modellkompetenz im Biologieunterricht Klasse 7-10* (pp. 58–66). Donauwörth: Auer.

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*(3), 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). 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. Part 6. Nature of science: History, philosophy and sociology of science.* Nicosia, Cyprus: European Science Education Research Association.

Krell, M., Upmeier zu Belzen, A., & Krüger, D. (2014b). How year 7 to year 10 students categorise models: moving towards a student-based typology of biological models. In M. Ekborg & D. Krüger (Eds.), *Research in biological education* (pp. 117–131). Westermann.

Krell, M., Upmeier zu Belzen, A., & Krüger, D. (2014c). 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. (2016). Modellkompetenz im Biologieunterricht. In A. Sandmann & P. Schmiemann (Eds.), *BIOLOGIE lernen und lehren: Vol. 1. Biologiedidaktische Forschung: Schwerpunkte und Forschungsstände* (pp. 83–102). Berlin: Logos.

Krosnick, J. A. (1999). Survey research. *Annual review of psychology*, *50*(1), 537–567.

Krüger, D., & Burmester, A. (2005). Wie Schüler Pflanzen ordnen. *Zeitschrift für Didaktik der Naturwissenschaften*, *11*, 85–102.

Krüger, D., Upmeier zu Belzen, A., & Krell, M. (2016). Diskussionsbeitrag: Zu Primärreaktionen der Fotosynthese. *Der mathematische und naturwissenschaftliche Unterricht*, *4*, 277–279.

Kuckartz, U. (2016). *Qualitative Inhaltsanalyse: Methoden, Praxis, Computerunterstützung* (3rd ed.). Weinheim: Beltz.

Kunnan, A. J. (2007). Test Fairness, Test Bias, and DIF. *Language Assessment Quarterly*, *4*(2), 109–112. https://doi.org/10.1080/15434300701375865

Laing, G. J. (1981). The Classification of Models A Proposal. *Interdisciplinary Science Reviews*, *6*(4), 355–363. https://doi.org/10.1179/isr.1981.6.4.355

Lange-Schubert, K., Gogolin, S., & Krell, M. (2016, October). *Modellkompetenz bei Grundschüler_innen [invited presentation]*. Fachtagung an der Pädagogischen Hochschule Salzburg Stefan Zweig, Salzburg.

Leach, J., Millar, R., Ryder, J., & Séré, M.-G. (2000). Epistemological understanding in science learning: the consistency of representations across contexts. *Learning and Instruction*, *10*(6), 497–527.

Leatherdale, W. H. (1974). *The role of analogy, model and metaphor in science*. Amsterdam: North-Holland.

Lederman, N. G. (1992). Students' and teachers' conceptions of the nature of science: A review of the research. *Journal of Research in Science Teaching*, *29*(4), 331–359.

Lederman, N. G., & Abell, S. K. (Eds.). (2014). *Handbook of research on science education*. New York, NY: Routledge.

Lederman, N. G., & Lederman, J. S. (2014). Research on Teaching and Learning of Nature of Science. In N. G. Lederman & S. K. Abell (Eds.), *Handbook of research on science education* (pp. 600–620). New York, NY: Routledge.

Lederman, N. G., Abd-El-Khalick, F., Bell, R. L., & Schwartz, R. S. (2002). Views of nature of science questionnaire: Toward valid and meaningful assessment of learners' conceptions of nature of science. *Journal of Research in Science Teaching*, *39*(6), 497–521.

Lee, S.-Y., Chang, H.-Y., & Wu, H.-K. (2015). Students' Views of Scientific Models and Modeling: Do Representational Characteristics of Models and Students' Educational Levels Matter? *Research in Science Education*, 1–24. 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*(4), 6–15. https://doi.org/10.1111/j.1745-3992.2004.tb00164.x

Leighton, J., & Gierl, M. (2007b). Defining and evaluating models of cognition used in educational measurement to make inferences about

examinees' thinking processes. *Educational Measurement: Issues and Practice*, *26*(2), 3–16.

Leighton, J., & Gierl, M. (Eds.). (2007a). *Cognitive Diagnostic Assessment for Education: Theory and Applications*. Cambridge: Cambridge University Press.

Leisner-Bodenthin, A. (2006). Zur Entwicklung von Modellkompetenz im Physikunterricht. *Zeitschrift für Didaktik der Naturwissenschaften*, *12*, 91–109.

Louca, L. T., & Zacharia, Z. C. (2012). Modeling-based learning in science education: Cognitive, metacognitive, social, material and epistemological contributions. *Educational Review*, *64*(4), 471–492. https://doi.org/10.1080/00131911.2011.628748

MacIver, R., Anderson, N., Costa, A., & Evers, A. (2014). Validity of Interpretation: A user validity perspective beyond the test score. *International Journal of Selection and Assessment*, *22*, 149–164. https://doi.org/10.1111/ijsa.12065

Magnani, L., Nersessian, N. J., & Thagard, P. (Eds.). (1999). *Model-Based Reasoning in Scientific Discovery*. Boston: Springer.

Mahr, B. (2008). Ein Modell des Modellseins: Ein Beitrag zur Aufklärung des Modellbegriffs. In U. Dirks & E. Knobloch (Eds.), *Modelle* (pp. 187–218). Frankfurt am Main: Peter Lang.

Mahr, B. (2009). Die Informatik und die Logik der Modelle. *Informatik-Spektrum*, *32*(3), 228–249.

Mahr, B. (2011). On the Epistemology of Models. *Rethinking Epistemology*, *1*, 301–352.

Mahr, B., & Wendler, R. (2009). Modelle als Akteure: Fallstudien. *KIT-Report*, *156*, 1–74. Retrieved from https://www.flp.tu-berlin.de/fileadmin/fg53/KIT-Reports/r156.pdf

Martinez, M. E. (1999). Cognition and the question of test item format. *Educational Psychologist, 34*(4), 207–218.

Masters, G. N. (1982). A rasch model for partial credit scoring. *Psychometrika, 47*(2), 149–174. https://doi.org/10.1007/BF02296272

Mathesius, S., & Gogolin, S. (in press). Die intraspezifische Konkurrenz. In T. Bruckermann & K. Schlüter (Eds.), *Forschendes Lernen im Labor. Eine praktische Anleitung für die Lehramtsausbildung Biologie.* Berlin: Springer Spektrum.

Mathesius, S., Hartmann, S., Upmeier zu Belzen, A., & Krüger, D. (2016). Scientific Reasoning as an Aspect of Pre-Service Biology Teacher Education: Assessing Competencies Using a Paper-Pencil Test. In T. Tal & A. Yarden (Eds.), *The Future of Biology Education Research* (pp. 93–110). Haifa, Israel: The Technion, Israel Institute of Technology.

Mayer, J. (2007). Erkenntnisgewinnung als wissenschaftliches Problemlösen. In D. Krüger & H. Vogt (Eds.), *Theorien in der biologiedidaktischen Forschung. Ein Handbuch für Lehramtsstudenten und Doktoranden* (pp. 177–186). Berlin: Springer.

McCloy, R. A., Heggestad, E. D., & Reeve, C. L. (2005). A Silk Purse From the Sow's Ear: Retrieving Normative Information From Multidimensional Forced-Choice Items. *Organizational Research Methods, 8*(2), 222–248. https://doi.org/10.1177/1094428105275374

McCoach, D. B., Gable, R. K., & Madura, J. P. (2013). Evidence Based on Relations to Other Variables: Bolstering the Empirical Validity Arguments for Constructs. In D. B. McCoach, R. K. Gable, & J. P. Madura (Eds.), *Instrument Development in the Affective Domain. School and Corporate Applications* (3rd ed.). New York, NY: Springer.

McComas, W. F. (2008). Seeking historical examples to illustrate key aspects of the nature of science. *Science & Education*, *17*(2-3), 249–263. https://doi.org/10.1007/s11191-007-9081-y

McComas, W. F. (2013). *The language of science education: An expanded glossary of key terms and concepts in science teaching and learning*. Rotterdam: Sense Publishers.

Meisert, A. (2007). Über den Umgang mit Hypothesen. *Der mathematische und naturwissenschaftliche Unterricht*, *60*, 431–437.

Messick, S. (1984). The psychology of educational measurement. *Journal of educational measurement*, *21*(3), 215–237.

Messick, S. (1995). Validity of psychological assessment: validation of inferences from persons' responses and performances as scientific inquiry into score meaning. *American psychologist*, *50*(9), 741–749.

Mittelstraß, J. (2004). *Enzyklopädie Philosophie und Wissenschaftstheorie*. Stuttgart: Metzler.

Moosbrugger, H., & Kelava, A. (2012). *Testtheorie und Fragebogenkonstruktion*. Berlin: Springer.

Morgan, M. S., & Morrison, M. (Eds.). (1999). *Ideas in context: Vol. 52. Models as mediators: Perspectives on natural and social science*. Cambridge: Cambridge University Press.

Morse, J. M. (2003). Principles of mixed methods and multimethod research design. In A. Tashakkori & C. Teddlie (Eds.), *Handbook of mixed methods in social & behavioral sciences* (pp. 189–208). Thousand Oaks: Sage.

Namdar, B., & Shen, J. (2015). Modeling-Oriented Assessment in K-12 Science Education: A synthesis of research from 1980 to 2013 and new directions. *International Journal of Science Education*, *37*, 993–1023.

Nehm, R. H., & Ha, M. (2011). Item feature effects in evolution assessment. *Journal of Research in Science Teaching*, *48*(3), 237–256.

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

Nichols, P. (1994). A Framework for Developing Cognitively Diagnostic Assessments. *Review of Educational Research*, *64*(4), 575–603. https://doi.org/10.3102/00346543064004575

Nicolaou, C., & 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

Nisbett, R. E., & Wilson, T. D. (1977). Telling more than we can know: Verbal reports on mental processes. *Psychological Review*, *84*(3), 231–259.

Nowak, K. H., Nehring, A., Tiemann, R., & Upmeier zu Belzen, A. (2013). Assessing students' abilities in processes of scientific inquiry in biology using a paper-and-pencil test. *Journal of Biological Education*, *47*(3), 182–188. https://doi.org/10.1080/00219266.2013.822747

NRC. (2001). *Knowing What Students Know*. Washington, DC: National Academies Press.

NRC. (2012). *A framework for K-12 science education: Practices, crosscutting concepts, and core ideas*. Washington, DC: National Academic Press.

Odenbaugh, J. (2005). Idealized, inaccurate but successful: A pragmatic approach to evaluating models in theoretical ecology. *Biology and Philosophy*, *20*(2), 231–255.

Oh, P. S., & Oh, S. J. (2011). What teachers of science need to know about models: An overview. *International Journal of Science Education*, *33*(8), 1109–1130. https://doi.org/10.1080/09500693.2010.502191

Orsenne, J. (2015). *Aktivierung von Schülervorstellungen zu Modellen durch praktische Tätigkeiten der Modellbildung*. Dissertation. Retrieved

from http://edoc.hu-berlin.de/dissertationen/orsenne-juliane-2015-11-26/PDF/orsenne.pdf

Passmore, C., Gouvea, J., & Giere, R. (2014). Models in science and in learning science. In M. Matthews (Ed.), *International handbook of research in history, philosophy and science teaching* (pp. 1171–1202). Dordrecht: Springer.

Patzke, C., Krüger, D., & Upmeier zu Belzen, A. (2015). Entwicklung von Modellkompetenz im Längsschnitt. In M. Hammann, J. Mayer, & N. Wellnitz (Eds.), *Lehr- und Lernforschung in der Biologiedidaktik* (6th ed., pp. 43–58). Innsbruck: Studienverlag.

Paulsen, M. B., & Wells, C. T. (1998). Domain differences in the epistemological beliefs of college students. *Research in higher education*, *39*(4), 365–384.

Pluta, W. J., Chinn, C. A., & Duncan, R. G. (2011). Learners' epistemic criteria for good scientific models. *Journal of Research in Science Teaching*, *48*(5), 486–511.

Prins, G. T., Bulte, A. M. W., & Pilot, A. (2011). Evaluation of a design principle for fostering students' epistemological views on models and modelling using authentic practices as contexts for learning in chemistry education. *International Journal of Science Education*, *33*(11), 1539–1569.

Raykov, T., & Marcoulides, G. A. (2011). *Introduction to psychometric theory*. New York: Routledge.

Raymond, M., Lane, S., & Haladyna, T. (Eds.). (2015). *Handbook of Test Development* (2nd ed.). New York: Routledge.

Rayner, K., Warren, T., Juhasz, B. J., & Liversedge, S. P. (2004). The effect of plausibility on eye movements in reading. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *30*(6), 1290–1301.

Rios, J., & Wells, C. (2014). Validity evidence based on internal structure. *Psicothema*, *26*(1), 108–116.

Ropohl, M., Walpuski, M., & Sumfleth, E. (2015). Welches Aufgabenformat ist das richtige? – Empirischer Vergleich zweier Aufgabenformate zur standardbasierten Kompetenzmessung. *Zeitschrift für Didaktik der Naturwissenschaften*, *21*(1), 1–15.

Ross, R. M., Duggan-Haas, D., & Allmon, W. D. (2013). The Posture of Tyrannosaurus rex: Why Do Student Views Lag Behind the Science? *Journal of Geoscience Education*, *61*(1), 145–160. https://doi.org/10.5408/11-259.1

Ruhrig, J., & Höttecke, D. (2015). Was, wenn das Experiment nicht klappt? Unsichere Evidenz als Lerngelegenheit nutzen. (Themenheft Experimentieren). *Unterricht Physik*, *144*.

Ruppert, W. (2011). Modellorganismen. *Unterricht Biologie*, *363*.

Sadovnik, A. R., O'Day, J. A., Bohrnstedt, G. W., & Borman, K. M. (2013). *No Child Left Behind and the reduction of the achievement gap: Sociological perspectives on federal educational policy*. New York: Routledge.

Sandmann, A. (2014). Lautes Denken - die Analyse von Denk-, Lern- und Problemlöseprozesse. In D. Krüger, I. Parchmann, & H. Schecker (Eds.), *Methoden in der naturwissenschaftsdidaktischen Forschung* (pp. 179–188). Heidelberg: Springer.

Schecker, H. (2012). Standards, Competencies and Outcomes: A Critical View. In S. Bernholt, K. Neumann, & P. Nentwig (Eds.), *Making it tangible. Learning outcomes in science education* (pp. 219–234). Münster: Waxmann.

Scheersoi, A., & Dierkes, P. (2012). Zeig mir deine Zähne, und ich sage dir was du frisst. *Unterricht Biologie*, *374*, 12–19.

Schmiemann, P. (2010). *Modellierung von Schülerkompetenzen im Bereich des biologischen Fachwissens.* Berlin: Logos.

Schmiemann, P., & Lücken, M. (2014). Validität – Misst mein Test, was er soll? In D. Krüger, I. Parchmann, & H. Schecker (Eds.), *Methoden in der naturwissenschaftsdidaktischen Forschung* (pp. 107–120). Heidelberg: Springer.

Schwartz, R., & Lederman, N. G. (2005). *What scientists say: Scientists' views of models*, Paper presented at the annual meeting of the National Association for Research in Science Teaching, Dallas, USA.

Schwarz, C. (2002). The Role Of Meta-Modeling Knowledge In Learning With Models. Retrieved from http://schwarz.wiki.educ.msu.edu/file/view/Schwarz_ICLS_metapaper.pdf

Schwarz, C. V., & Gwekwerere, Y. N. (2007). Using a guided inquiry and modeling instructional framework (EIMA) to support preservice K-8 science teaching. *Science Education*, *91*(1), 158–186.

Schwarz, C. V., & White, B. Y. (2005). Metamodeling Knowledge: Developing Students' Understanding of Scientific Modeling. *Cognition and Instruction*, *23*(2), 165–205. https://doi.org/10.1207/s1532690xci2302_1

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

Schwarz, C., Reiser, B. J., Acher, A., Kenyon, L., & Fortus, D. (2012). MoDeLS: Challenges in defining a learning progression for scientific modeling. In A. C. Alonzo & A. W. Gotwals (Eds.), *Learning Progressions in Science. Current Challenges and Future Directions* (pp. 101–137). Rotterdam: SensePublishers.

SenBJF (Ed.). (2015). *Berliner Rahmenlehrplan – Biologie Jahrgangsstufen 7-10*: Retrieved from: http://bildungsserver.berlin-brandenburg.de/rlp-online/c-faecher/biologie/kompetenzentwicklung.

Shepard, L. A. (1997). The centrality of test use and consequences for test validity. *Educational Measurement: Issues and Practice*, *16*(2), 5–24.

Sins, P., Savelsbergh, E., van Joolingen, W., & 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*(9), 1205–1229.

SMI SensoMotoric Instruments. (2016). BeGaze Manual Version 3.6. Retrieved from http://www.smivision.com/en/gaze-and-eye-tracking-systems/support/software-download.html

Smit, J. J. A., & Finegold, M. (1995). Models in physics: Perceptions held by final-year prospective physical science teachers studying at South African universities. *International Journal of Science Education*, *17*(5), 621–634. https://doi.org/10.1080/0950069950170506

Snow, R. E., & Lohman, D. F. (1989). *Implications of cognitive psychology for educational measurement*. New York: Macmillan.

Sophian, C. (1997). Beyond competence: The significance of performance for conceptual development. *Cognitive Development*, *12*(3), 281–303. https://doi.org/10.1016/S0885-2014(97)90001-0

Spearman, C. (1904). The proof and measurement of association between two things. *The American journal of psychology*, *15*(1), 72–101.

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

Strike, K. A., & Posner, G. J. (1992). A revisionist theory of conceptual change. In R. A. Duschl & R. J. Hamilton (Eds.), *SUNY series in science education. Philosophy of science, cognitive psychology, and educational theory and practice* (pp. 147–176). New York: State University Press.

Suckling, C., Suckling, K., & Suckling, C. (1978). *Chemistry through models*. Cambridge: Cambridge University Press.

Suppe, F. (Ed.). (1974). *The structure of scientific theories*. [Symposium on the Structure of Scientific Theories held in Urbana, March 26 to 29, 1969]. Urbana: University of Illinois Press.

Suppes, P. (1961). A Comparison of the Meaning and Uses of Models in Mathematics and the Empirical Sciences. In H. Freudenthal (Ed.), *The Concept and the Role of the Model in Mathematics and Natural and Social Sciences* (pp. 163–177). Dordrecht: Springer. https://doi.org/10.1007/978-94-010-3667-2_16

Tarski, A. (1966). *Einführung in die mathematische Logik*. Göttingen: Vandenhoeck & Ruprecht.

Terzer, E. (2012). *Modellkompetenz im Kontext Biologieunterricht – Empirische Beschreibung von Modellkompetenz mithilfe von Multiple-Choice Items.* Dissertation. Retrieved from http://edoc.hu-berlin.de/dissertationen/terzer-eva-2012-12-19/PDF/terzer.pdf

Treagust, D., Chittleborough, G., & Mamiala, T. (2002). Students' understanding of the role of scientific models in learning science. *Journal of Science Education*, *24*(4), 357–368. https://doi.org/10.1080/09500690110066485

Treagust, D., Chittleborough, G., & Mamiala, T. (2004). Students' Understanding of the Descriptive and Predictive Nature of Teaching Models in Organic Chemistry. *Research in Science Education*, *34*(1), 1–20. https://doi.org/10.1023/B:RISE.0000020885.41497.ed

Trier, U., Krüger, D., & Upmeier zu Belzen, A. (2014). Students' versus Scientists' Conceptions of Models and Modelling. In M. Ekborg, D. Krüger, D. J. Boerwinkel, M. Ergazaki, M. J. Gil Quilez, G. Molinatti et al. (Ed.), *Research in Biological Education* (pp. 103–115). Berlin.

Ubben, I., Nitz, S., Rousseau, M., & Upmeier zu Belzen, A. (2015). Modelle von und für Evolution in Schulbüchern. In U. Gebhard, M. Hammann, & B. Knälmann (Eds.), *Bildung durch Biologieunterricht. 20. Internationale Tagung der Fachsektion Didaktik der Biologie (FDdB) im VBiO* (pp. 75–76). Retrieved from http://www.biodidaktik.de/upload/downloads/1443164473.pdf

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

van der Valk, T., van Driel, J. H., & Vos, W. de. (2007). Common Characteristics of Models in Present-day Scientific Practice. *Research in Science Education*, *37*(4), 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*(11), 1141–1153. https://doi.org/10.1080/095006999290110

van Driel, J. H., & Verloop, N. (2002). Experienced teachers' knowledge of teaching and learning of models and modelling in science education. *International Journal of Science Education*, *24*(12), 1255–1272.

van Heteren-Frese, C. (2016). *Untersuchung von Diagnoseaufgaben zum Modellverstehen durch Eye-Tracking* (Masterarbeit). Freie Universität Berlin.

van Joolingen, W. (2004). *Roles of modeling in inquiry learning.* Paper presented at IEEE International Conference.

van Oers, B. (1998). From context to contextualizing. *Learning and Instruction, 8*(6), 473–488. https://doi.org/10.1016/S0959-4752(98)00031-0

Verhoeff, R. P., Waarlo, A. J., & Boersma, K. T. (2008). Systems modelling and the development of coherent understanding of cell biology. *International Journal of Science Education*, *30*(4), 543–568.

Vo, T., Forbes, C. T., Zangori, L., & Schwarz, C. V. (2015). Fostering Third-Grade Students' Use of Scientific Models with the Water Cycle: Elementary teachers' conceptions and practices. *International Journal of Science Education*, *37*(15), 2411–2432. https://doi.org/10.1080/09500693.2015.1080880

Warm, T. A. (1989). Weighted likelihood estimation of ability in item response theory. *Psychometrika*, *54*(3), 427–450.

Webb, N. M., Shavelson, R. J., & Haertel, E. H. (2007). Reliability Coefficients and Generalizability Theory. In C. R. Rao (Ed.), *Handbook of Statistics* (pp. 81–124). Amsterdam: Elsevier. https://doi.org/10.1016/S0169-7161(06)26004-8

Weinert, F. E. (2001). *Leistungsmessungen in Schulen*. Weinheim: Beltz.

Wellnitz, N. (2012). *Kompetenzstruktur und -niveaus von Methoden naturwissenschaftlicher Erkenntnisgewinnung*. Berlin: Logos.

Williams, G., & Clement, J. (2015). Identifying multiple levels of discussion-based teaching strategies for constructing scientific models. *International Journal of Science Education*, *37*(1), 82–107.

Windschitl, M., & Thompson, J. (2006). Transcending Simple Forms of School Science Investigation: The Impact of Preservice Instruction on Teachers' Understandings of Model-Based Inquiry. *American Educational Research Journal*, *43*(4), 783–835. https://doi.org/10.3102/00028312043004783

Wirtz, M., & Caspar, F. (2002). *Beurteilerübereinstimmung und Beurteilerreliabilität: Methoden zur Bestimmung und Verbesserung der*

Zuverlässigkeit von Einschätzungen mittels Kategoriensystemen und Ratingskalen. Göttingen: Hogrefe.

Wu, M., Adams, R., Wilson, & Haldane, S. (2007). *ACER ConQuest*. Camberwell, VIC: ACER Press.

Yauch, C. A., & Steudel, H. J. (2003). Complementary Use of Qualitative and Quantitative Cultural Assessment Methods. *Organizational Research Methods, 6*(4), 465–481. https://doi.org/10.1177/1094428103257362

Zabel, J. (2001). DNA – ein interessantes Spielzeug? Unterrichtsanregung für die Sekundarstufe II. *Unterricht Biologie, 268*, 37–43.

9. Appendix

9.1. Publications

Publications in journals and handbooks

Gogolin, S. & Krüger, D. (in press). Students' understanding of the nature and purpose of models. *Journal of Research in Science Teaching*.

Gogolin, S., & Krüger, D. (in press). Modellverstehen im Biologieunterricht diagnostizieren und fördern. *Der mathematische und naturwissenschaftliche Unterricht.*

Böschl, F., Gogolin, S., Lange-Schubert, K., & Hartinger, A. (in press). Modellverstehen von Grundschülerinnen und Grundschülern in Abhängigkeit von Kontext und Kompetenzniveau. In Jahresband GDSU.

Reinisch, B, Krell, M., Hergert, S., Gogolin, S., & Krüger, D. (2017). Methodical challenges concerning the Draw-A-Scientist Test: A critical view about the assessment and evaluation of learners' conceptions of scientists. *International Journal of Science Education*, doi:10.1080/09500693.2017.1362712

Gogolin, S., Krell, M., Lange-Schubert, K., Hartinger, A., Upmeier zu Belzen, A., & Krüger, D. (2017). Erfassung von Modellkompetenz bei Grundschüler_innen. In H. Giest, A. Hartinger, & S. Tänzer (Eds.), *Vielperspektivität im Sachunterricht* (pp. 108–115). Bad Heilbrunn: Klinkhardt-Verlag.

Mathesius, S., & Gogolin, S. (2017). Die Letzten werden die Ersten sein – Praktisches Modellieren zu Körperformen von Plankton. *Biologie 5-10, 17*, 10-13.

Mathesius, S., & Gogolin, S. (2017). Die intraspezifische Konkurrenz. In T. Bruckermann & K. Schlüter (Eds.), *Forschendes Lernen im Labor. Eine praktische Anleitung für die Lehramtsausbildung Biologie.* Berlin: Springer Spektrum.

Gogolin, S., & Krüger, D. (2016). Konstruktion von Diagnoseaufgaben zum Zweck von Modellen. *Biologie Lehren und Lernen – Zeitschrift für Didaktik der Biologie, 1*(20), 44–62.

Gogolin, S. & Krüger, D. (2016). Diagnosing Students' Understanding of the Nature of Models. *Research in Science Education, 47, 5,* 1127–1149, doi: 10.1007/s11165-016-9551-9

Weitzel, H., Gogolin, S. & Mathesius, S. (2016). Arielle – Fisch oder Säugetier? *Unterricht Biologie, 413,* 9-14.

Gogolin, S. & Krüger, D. (2016). Diagnosing Students' Understanding of Models and Modeling [Abstract]. In M. Atwater, M.-H. Chiu, W. Kyle, T. Sondergeld, & R. Yerrick (Hrsg.), *NARST Annual International Conference Conference Programm Book. Toward Equity and Justice.* (p. 349). Baltimore.

Gogolin, S. (2015). Korallenbleiche – Eine Symbiose in Gefahr. *Unterricht Biologie, 407,* 12-17.

Gogolin, S., & Krüger, D. (2015). Nature of models - Entwicklung von Diagnoseaufgaben. In M. Hammann, J. Mayer, & N. Wellnitz (Eds.), *Lehr- und Lernforschung in der Biologiedidaktik* (6th ed., pp. 27–41). Innsbruck: Studienverlag.

Gogolin, S. & Krüger, D. (2015). Einsatz unterschiedlicher Befragungsformate zur Validierung eines Diagnoseinstruments zum Modellverstehen [Abstract]. In U. Gebhard, M. Hammann, & B. Knälmann (Hrsg.), *Bildung durch Biologieunterricht. 20. Internationale Tagung der Fachsektion Didaktik der Biologie (FDdB) im VBiO* (S. 73-74). Hamburg.

Gogolin, S. & Genzel, F. (2015). Ab in die Pilze - Eng verbunden. *Unterricht Biologie, 406,* 8.

Gogolin, S. & Mathesius, S. (2014). Königliche Bienen. *Unterricht Biologie, 400,* 52-54.

Gogolin, S. & Mathesius, S. (2014). Gleich und gleich gesellt sich gern - oder nicht? *Unterricht Biologie, 394,* 21-25.

Gogolin, S. & Krüger, D. (2014). Forced Choice-Aufgaben zum Modellverstehen. Validierungsstudie mit lautem Denken [Abstract]. In L. Kotzebue, A. Dittmer, A. Möller, & P. Schmiemann (Hrsg.), *17. Internationale Frühjahrsschule der Fachsektion Didaktik der Biologie im VBIO* (S. 19-20).

Gogolin, S. & Krüger, D. (2014). Modellverstehen im Biologieunterricht - Evaluation einer Diagnosestrategie [Abstract]. In D. Chernyak, A. Möller, A. Dittmer, & Schmiemann, P. (Hrsg.), *16. Internationale Frühjahrsschule der Fachsektion Didaktik der Biologie im VBIO* (S. 110-111).

Gogolin, S. & Krüger, D. (2013). Diagnose von Modellkompetenz - Entwicklung eines Instruments auf der Basis eines Kategoriensystems aus Schüleraussagen [Abstract]. In J. Mayer, M. Hammann, N. Wellnitz, J. Arnold, & M. Werner (Hrsg.), *Theorie, Empirie, Praxis: 19. Internationale Tagung der Fachsektion Didaktik der Biologie (FDdB) im VBiO* (S. 56-57). Kassel: Universität Kassel.

Posters and papers personally presented at conferences

Lange-Schubert, K., Gogolin, S. & Krell, M. (2016, Oktober). *Modellkompetenz bei Grundschüler_innen* [Eingeladener Vortrag]. Fachtagung an der Pädagogischen Hochschule Salzburg Stefan Zweig, 26.-27.10.2016, Salzburg.

Gogolin, S. (2016, Mai). *Entwicklung eines Instruments zur Erfassung von Modellierungskompetenz bei Grundschüler_innen* [Eingeladener Vortrag]. Kolloquium der Pädagogischen Psychologie mit Schwerpunkt Lehren, Lernen und Entwicklung. 31.05.2016. Leipzig.

Gogolin, S. & Krüger, D. (2016, April). *Diagnosing Students' Understanding of Models and Modeling* [Vortrag]. Annual International Conference of the National Association for Research in Science Teaching (NARST), 14.-17.04.2016, Baltimore.

Lange-Schubert, K., Gogolin, S., Krell, M., Krüger, D., Upmeier zu Belzen, A. & Hartinger, A. (2016, März). *Erfassung von Modellierungskompetenz bei Grundschüler_innen* [Vortrag]. 25. Jahrestagung der Gesellschaft für Didaktik des Sachunterrichts (GDSU), 03.-05.03.2016, Erfurt.

Gogolin, S. & Krüger, D. (2015, September). Einsatz unterschiedlicher Befragungsformate zur Validierung eines Diagnoseinstruments zum Modellverstehen [Vortrag]. 20. Internationale Tagung der Fachsektion Didaktik der Biologie (FDdB) im VBiO, 14.-17.09.2015, Hamburg.

Gogolin, S. & Krüger, D. (2015, September). *Students' understanding of models in biology: Empirical validation of a diagnostic instrument* [Vortrag]. 11th Conference of the European Science Education Research Association (ESERA), 31.08.-04.09.2015, Helsinki.

Gogolin, S. & Krüger, D. (2015, Februar). Forced Choice-Aufgaben zum Modellverstehen. Validierungsstudie mit lautem Denken [Vortrag]. 17. Internationale Frühjahrsschule der Fachsektion Didaktik der Biologie im VBiO, 23.-26.02.2015, München.

Gogolin, S. & Krüger, D. (2014, Februar). *Modellverstehen im Biologieunterricht - Evaluation von Diagnosestrategien* [Poster]. 16. Internationale Frühjahrsschule der Fachsektion Didaktik der Biologie (FDdB) im VBiO, 24.-27.02.2014, Trier.

Gogolin, S. & Krüger, D. (2013, September). *Diagnose von Modellkompetenz: Entwicklung eines Instruments auf der Basis eines Kategoriensystems aus Schüleraussagen* [Poster]. 19. Internationale Tagung der Fachsektion Didaktik der Biologie (FDdB) im VBiO, 16.-20.09.2013, Kassel.

9.2. Abbreviations[49]

AERA	American Educational Research Association
APA	American Psychological Association
EAP/PV	Expected a Posteriori / Plausible Values
IRT	Item-Response-Theorie
KMK	Ständige Konferenz der Kultusminister der Länder in der Bundesrepublik
n	Partial quantity
N	Total quantity
NCME	National Council on Measurement in Education
NGSS	Next generation science standards
NRC	National Research Council
PISA	Programme for International Student Assessment
SenBJF	Senatsverwaltung für Bildung, Jugend und Familie
WLE	Weighted Likelihood Estimator

Contexts:

AR	Archaeopteryx
AS	Air stream
BG	Bacterial growth
BM	Bio membrane
BR	Brain
EV	Evolution
JF	Jura forest
LZ	Lakeshore zone
NT	Neanderthal man
TR	*Tyrannosaurus rex*
VS	Influenza virus
WC	Water circle
WM	Water melon

[49] This list only contains abbreviations that have been used on more than one occasion in which they were not explanined again.

9.3. Tables

9.4. Figures

9.5. Versions of answer options

Version 1 of the answer options in the aspect ‘nature of models’ ($N1_9$)

Modell	T. rex	Neandertaler	Archaeopteryx	Biomembran	Virus	Evolution
N I	… sieht so aus wie ein damals lebender *Tyrannosaurus rex*, weil das Modell eine maßstabsgetreue Nachbildung des Knochenfundes ist.	… sieht so aus wie ein damals lebender Neandertaler, weil das Modell eine maßstabsgetreue Nachbildung des Schädelfundes ist.	… sieht so aus wie ein damals lebender Urvogel, weil das Modell eine maßstabsgetreue Nachbildung des Fossilienfundes ist.	… sieht so aus wie eine echte Biomembran, weil das Modell eine vergrößerte Nachbildung von mikroskopischen Aufnahmen ist.	… sieht so aus wie ein echter Grippe-Virus, weil das Modell eine vergrößerte Nachbildung von mikroskopischen Aufnahmen ist.	… stellt die Evolution des Menschen richtig dar, weil das Modell eine Zusammenstellung der wichtigsten Schädelfunde zeigt.
	… ähnelt dem damals lebenden *Tyrannosaurus rex* sehr, weil viele Wissenschaftler an der Entwicklung des Modells gearbeitet haben.	… ähnelt dem damals lebenden Neandertaler sehr, weil viele Wissenschaftler an der Entwicklung des Modells gearbeitet haben.	… ähnelt dem damals lebenden Urvogel sehr, weil viele Wissenschaftler an der Entwicklung des Modells gearbeitet haben.	… ähnelt einer echten Biomembran sehr, weil viele Wissenschaftler an der Herstellung der Zeichnung gearbeitet haben.	… ähnelt einem echten Grippe-Virus sehr, weil viele Wissenschaftler an der Entwicklung des Modells gearbeitet haben.	… stellt die Evolution des Menschen richtig dar, einige Eigenschaften des Modells, z. B. die farbigen Äste, sind jedoch unnötig.
	… sieht nicht so aus wie ein damals lebender *Tyrannosaurus rex*, weil er meiner Meinung nach anders aussah.	… sieht so aus wie ein damals lebender Neandertaler, weil ich mir z. B. die Augenwülste bzw. die Nase ungefähr so vorstelle.	… sieht so aus wie ein damals lebender Urvogel, weil ich mir z. B. die Flügel bzw. den Schwanz ungefähr so vorstelle.	… sieht so aus wie eine echte Biomembran, weil ich mir z. B. den Aufbau aus zwei Schichten ungefähr so vorstelle.	… sieht so aus wie ein echter Grippe-Virus, weil ich mir z. B. die Form bzw. die Oberfläche ungefähr so vorstelle.	… zeigt die Evolution des Menschen so, wie ich sie mir z. B. durch das Fernsehen oder den Biologieunterricht vorstelle.
N II	… gleicht z. B. in Bezug auf die kurzen Vorderbeine dem damals lebenden *Tyrannosaurus rex*, über andere Merkmale kann man aber nichts aussagen.	… zeigt nur wesentliche Eigenschaften des damals lebenden Neandertalers, z. B. die starken Augenwülste oder die breite Nase.	… gleicht z. B. in Bezug auf die Schwanzfedern dem damals lebenden Urvogel, über andere Merkmale kann man aber nichts aussagen.	… gleicht z. B. in Bezug auf die Anordnung der Teilchen einer echten Biomembran, über andere Merkmale kann man aber nichts aussagen.	… gleicht z. B. in Bezug auf die Form einem Grippe-Virus, trotzdem kann ein echter Grippe-Virus auch anders aussehen.	…zeigt nur wesentliche Schritte der Evolution des Menschen, Zwischenformen werden jedoch nicht dargestellt.
	… zeigt nur wesentliche Eigenschaften des damals lebenden *Tyrannosaurus rex*, z. B. die kurzen Arme, das große Gebiss und den langen Schwanz.	… gleicht z. B. in Bezug auf seine Kopfform dem damals lebenden Neandertaler, trotzdem kann der echte Neandertaler auch anders ausgesehen haben.	… gleicht z. B. in Bezug auf seine Kopfform dem damals lebenden Urvogel, trotzdem kann der echte Urvogel auch anders ausgesehen haben.	… gleicht z. B. in Bezug auf die Form einer Biomembran, trotzdem kann eine echte Biomembran auch anders aussehen.	… zeigt nur wesentliche Eigenschaften eines echten Grippe-Virus, z. B. die runde Form oder die Strukturen auf der Oberfläche.	… stellt die Evolution des Menschen nachvollziehbar dar, trotzdem kann sie auch anders verlaufen sein.
	… gleicht z. B. in Bezug auf seine Form dem damals lebenden *Tyrannosaurus rex*, trotzdem kann der echte *Tyrannosaurus rex* auch anders ausgesehen haben.	… gleicht z. B. in Bezug auf die Augenwülste dem damals lebenden Neandertaler, über andere Merkmale kann man aber nichts aussagen.	… zeigt nur wesentliche Eigenschaften des damals lebenden Urvogels, z. B. die großen Schwanzfedern oder den schmalen Kopf.	… zeigt nur wesentliche Eigenschaften einer echten Biomembran, z. B. die Form oder die Anordnung der Teilchen.	… gleicht z. B. in Bezug auf die Form und die Oberfläche einem echten Grippe-Virus, über andere Merkmale kann man aber nichts aussagen.	…stellt die Evolution des Menschen teilweise richtig dar, über weitere, noch nicht entdeckte Arten sagt das Modell aber nichts aus.
N III	... zeigt den damals lebenden *Tyrannosaurus rex* so, wie Wissenschaftler den Knochenfund interpretieren.	... zeigt den damals lebenden Neandertaler so, wie Wissenschaftler den Schädelfund interpretieren.	… ähnelt dem damals lebenden Urvogel möglicherweise, Wissenschaftler können aber nur vermuten und nicht wissen wie er ausgesehen hat.	… ähnelt der echten Biomembran möglicherweise, Wissenschaftler können aber nur vermuten und nicht wissen wie sie aussieht.	... zeigt einen Grippe-Virus so, wie Wissenschaftler die mikroskopischen Aufnahmen interpretieren.	… zeigt die Evolution des Menschen so, wie es Wissenschaftler ausgehend von den Schädelfunden vermuten.
	… zeigt den damals lebenden *Tyrannosaurus rex* so, wie es Wissenschaftler ausgehend vom Knochenfund vermuten.	… zeigt den damals lebenden Neandertaler so, wie es Wissenschaftler ausgehend vom Schädelfund vermuten.	… zeigt den damals lebenden Urvogel so, wie es Wissenschaftler ausgehend vom Fossilienfund vermuten.	… zeigt die Biomembran so, wie es Wissenschaftler ausgehend von mikroskopischen Aufnahmen vermuten.	… zeigt einen Grippe-Virus so, wie es Wissenschaftler ausgehend von mikroskopischen Aufnahmen vermuten.	... zeigt die Evolution des Menschen so, wie Wissenschaftler sie sich ausgehend von den Schädelfunden vorstellen.
	… ähnelt dem damals lebenden *Tyrannosaurus rex* möglicher-weise, Wissenschaftler können aber nur vermuten und nicht wissen wie er ausgesehen hat.	… ähnelt dem damals lebenden Neandertaler möglicherweise, Wissenschaftler können aber nur vermuten und nicht wissen wie er ausgesehen hat.	... zeigt den damals lebenden Urvogel so, wie Wissenschaftler den Fossilienfund interpretieren.	... zeigt eine Biomembran so, wie Wissenschaftler die mikroskopischen Aufnahmen interpretieren.	… ähnelt einem echten Grippe-Virus möglicherweise, Wissenschaftler können aber nur vermuten und nicht wissen wie er aussieht.	… stellt die Evolution des Menschen möglicherweise richtig dar, Wissenschaftler können aber nur vermuten wie sie verlaufen ist.

Version 2 of the answer options in the aspect ‘nature of models’ ($N2_{6-9}$)

Modell	T. rex	Neandertaler	Biomembran	Virus	Evolution	Wasserkreislauf	Luftstrom	Wassermelone
N I	… stimmt mit dem damals lebenden T. rex überein, weil das Modell eine Nachbildung des Knochenfundes ist.	… stimmt mit dem damals lebenden Neandertaler überein, weil das Modell eine Nachbildung des Schädelfundes ist.	… sieht so aus wie eine vergrößerte Biomembran, weil das Modell z. B. die dicken Außenwände und den Spalt in der Mitte richtig darstellt.	… sieht so aus wie ein echter Grippe-Virus, weil das Modell z. B. die runde Form und die Strukturen auf der Oberfläche richtig darstellt.	… stellt die Evolution des Menschen so dar, wie sie verlaufen ist, weil Wissenschaftler in dem Modell die wichtigsten Schädelfunde zusammen-gestellt haben.	… stellt den Wasserkreislauf richtig dar, weil Wissenschaftler in dem Modell viele Wetteraufzeichnungen zusammengefasst haben.	… entspricht der tatsächlichen Luftmenge in einer Lunge, weil das Modell die Stärke und die Häufigkeit des Ein- und Ausatmens berücksichtigt.	… entspricht dem Gewicht einer echten Wassermelone, weil das Modell das Anfangsgewicht, den Wachstumsfaktor und die Zeit berücksichtigt.
	… gleicht dem damals lebenden T. rex, weil viele Wissenschaftler an der Entwicklung des Modells gearbeitet haben.	… gleicht dem damals lebenden Neandertaler, weil viele Wissenschaftler an der Entwicklung des Modells gearbeitet haben.	… bildet die echte Biomembran so ab, wie sie aussieht, weil es von Wissen-schaftlern mit Hilfe der neusten Technik hergestellt wurde.	… bildet einen echten Grippe-Virus so ab, wie er aussieht, weil es von Wissen-schaftlern mit Hilfe der neusten Technik hergestellt wurde.	… zeigt die Evolution des Menschen so, wie sie verlaufen ist, einige Eigenschaften des Modells, z. B. die farbigen Äste, sind jedoch unnötig.	… zeigt den Wasserkreislauf insgesamt richtig, einige Details des Modells, z. B. die Farben oder die Formen, sind jedoch unnatürlich.	… stimmt mit der tatsächlichen Luftmenge in einer Lunge überein, weil das Modell auf der Basis vieler Messungen entwickelt wurde.	… stimmt mit dem Gewicht einer echten Wassermelone überein, weil das Modell auf der Basis vieler Messungen entwickelt wurde.
	… zeigt, dass der damals lebende T. rex kurze Arme, ein großes Gebiss und einen langen Schwanz hatte.	… zeigt, dass der Neandertaler dicke Augenwülste, eine breite Nase und ein fliehendes Kinn hatte.	… stellt die echte Biomembran richtig dar, weil ich mir z. B. ihren Aufbau aus zwei Schichten so vorstelle.	… stellt den echten Grippe-Virus richtig dar, weil ich mir z. B. seine Form oder die Strukturen auf der Oberfläche so vorstelle.	-------	-------	-------	-------
N II	… gleicht in einigen Merkmalen dem damals lebenden T. rex, über andere Merkmale wissen die Wissenschaftler aber nichts.	… gleicht in einigen Merkmalen dem damals lebenden Neandertaler, über andere Merkmale wissen die Wissenschaftler aber nichts.	… bildet die echte Biomembran teilweise so ab, wie sie aussieht, über einige Merkmale haben die Wissenschaftler aber keine Informationen.	… bildet den echten Grippe-Virus teilweise so ab, wie er aussieht, über einige Merkmale haben die Wissenschaftler aber keine Informationen.	… stellt die Evolution des Menschen teilweise richtig dar, über weitere, noch nicht entdeckte Arten haben die Wissenschaftler aber keine Informationen.	… stellt den Wasserkreislauf richtig dar, trotzdem kann der Wasserkreislauf je nach Wetter auch anders verlaufen.	… stimmt mit der tatsächlichen Luftmenge in einer Lunge annähernd überein, das Modell trifft aber nicht auf jeden Menschen zu.	… stimmt mit dem Gewicht einer echten Wassermelone annähernd überein, Wassermelonen wachsen in der Natur aber nicht einheitlich.
	… stimmt z. B. in Bezug auf seine Körperform mit einem damals lebenden T. rex überein, trotzdem kann ein echter T. rex auch anders ausgesehen haben.	… stimmt z. B. in Bezug auf sein Kinn mit einem damals lebenden Neandertaler überein, trotzdem kann ein echter Neandertaler auch anders ausgesehen haben.	… sieht z. B. in Bezug auf den Aufbau so aus wie eine echte Biomembran, trotzdem kann es in der Natur auch Abweichungen von diesem Regelfall geben.	… sieht z. B. in Bezug auf seine Form so aus wie ein echter Grippe-Virus, trotzdem kann es in der Natur auch Grippe-Viren geben, die anders aussehen.	… zeigt nur einen Ausschnitt der Evolution des Menschen, weitere Schädel von früheren Vorfahren werden im Modell jedoch nicht dargestellt.	… zeigt nur wesentliche Schritte des Wasserkreislaufs, weitere Zwischenstationen des Wassers stellen die Wissenschaftler im Modell jedoch nicht dar.	... entspricht nur ungefähr der tatsächlichen Luftmenge in einer Lunge, denn das Modell bezieht nicht alle Faktoren beim Atmen in die Berechnung mit ein.	... entspricht nur ungefähr dem Gewicht einer echten Wassermelone, denn auch Umwelteinflüsse haben einen Einfluss auf deren Wachstum.
	… zeigt nur wesentliche Eigenschaften des Knochenfundes, z. B. die kurzen Arme, das große Gebiss und den langen Schwanz.	… zeigt nur wesentliche Eigenschaften des Schädelfundes, z. B. die starken Augenwülste, die breite Nase und das fliehende Kinn.	… stellt nur einen Ausschnitt der echten Biomembran dar, z. B. ihren Aufbau aus zwei Schichten oder die Anordnung der Teilchen.	… stellt nur einige Merkmale eines echten Grippe-Virus dar, z. B. die äußere Form oder die Strukturen auf der Oberfläche.	-------	-------	-------	-------
N III	... gleicht dem damals lebenden T. rex möglicherweise, sicher kann man aber nicht wissen, wie er ausgesehen hat.	... gleicht dem damals lebenden Neandertaler möglicherweise, sicher kann man aber nicht wissen, wie er ausgesehen hat.	... bildet eine Biomembran so ab, wie sie vielleicht aussieht, wie sie tatsächlich aussieht, kann man aber nicht wissen.	... bildet einen Grippe-Virus so ab, wie er vielleicht aussieht, wie er tatsächlich aussieht, kann man aber nicht wissen.	… stellt die Evolution des Menschen vielleicht richtig dar, sicher kann man aber nicht wissen, wie sie tatsächlich verlaufen ist.	… stellt den Wasserkreislauf vielleicht richtig dar, sicher kann man aber nicht wissen, wie er tatsächlich verläuft.	… entspricht der tatsächlichen Luftmenge in einer Lunge möglicherweise, die tatsächliche Menge an Luft kann das Modell aber nicht bestimmen.	… entspricht dem Gewicht einer echten Wassermelone möglicherweise, das tatsächliche Gewicht kann das Modell aber nicht bestimmen.
	… stimmt vielleicht mit dem damals lebenden T. rex überein, die Wissenschaftler können aber nur vermuten, wie er ausgesehen hat.	… stimmt vielleicht mit dem damals lebenden Neandertaler überein, die Wissenschaftler können aber nur vermuten, wie er ausgesehen hat.	… stellt die Biomembran so dar, wie es Wissenschaftler ausgehend von mikroskopischen Aufnahmen vermuten.	… stellt einen Grippe-Virus so dar, wie es Wissenschaftler ausgehend von mikroskopischen Aufnahmen vermuten.	… zeigt die Evolution des Menschen so, wie es Wissenschaftler ausgehend von den Schädelfunden vermuten.	… zeigt den Wasserkreislauf so, wie es Wissenschaftler ausgehend von Wetteraufzeichnungen vermuten.	… stimmt mit der tatsächlichen Luftmenge in einer Lunge vielleicht überein,man kann aber mit dem Modell nur vermuten, wie viel Luft in der Lunge ist.	… stimmt mit dem Gewicht einer echten Wassermelone vielleicht überein, man kann aber mit dem Modell nur vermuten, wie schwer diese ist.
	… zeigt den damals lebenden T. rex so, wie man es ausgehend vom Knochenfund vermuten kann.	… zeigt den damals lebenden Neandertaler so, wie man es ausgehend v vermuten kann.	… sieht möglicherweise so aus wie eine echte Biomembran, man kann aber nur vermuten, wie sie aussieht.	… sieht möglicherweise so aus wie ein echter Grippe-Virus, man kann aber nur vermuten, wie er aussieht.	-------	-------	-------	-------

Version 3 of the answer options in the aspect ‘nature of models’ ($N3_6$)

Modell	T. rex	Neandertaler	Biomembran	Virus	Evolution	Wasserkreislauf	Luftstrom	Wassermelone
N I	… stimmt mit dem damals lebenden T. rex überein, weil das Modell eine Nachbildung des Knochenfundes ist.	… gleicht dem damals lebenden Neandertaler, weil viele Wissenschaftler an der Entwicklung des Modells gearbeitet haben.	… sieht so aus wie eine vergrößerte Biomembran, weil das Modell z. B. die dicken Außenwände und den Spalt in der Mitte richtig darstellt.	… bildet einen echten Grippe-Virus so ab, wie er aussieht, weil es von Wissenschaftlern mit Hilfe der neusten Technik hergestellt wurde.	… stellt die Evolution des Menschen so dar, wie sie verlaufen ist, weil Wissenschaftler in dem Modell die wichtigsten Schädelfunde zusammengestellt haben.	… stellt den Wasserkreislauf richtig dar, weil Wissenschaftler in dem Modell viele Wetteraufzeichnungen zusammengefasst haben.	… entspricht der tatsächlichen Luftmenge in einer Lunge, weil das Modell die Stärke und die Häufigkeit des Ein- und Ausatmens berücksichtigt.	… entspricht dem Gewicht einer echten Wassermelone, weil das Modell das Anfangsgewicht, den Wachstumsfaktor und die Zeit berücksichtigt.
	… zeigt, dass der damals lebende T. rex kurze Arme, ein großes Gebiss und einen langen Schwanz hatte.	… zeigt, dass der Neandertaler dicke Augenwülste, eine breite Nase und ein fliehendes Kinn hatte.	… stellt die echte Biomembran richtig dar, weil ich mir z. B. ihren Aufbau aus zwei Schichten so vorstelle.	… sieht so aus wie ein echter Grippe-Virus, weil das Modell z. B. die runde Form und die Strukturen auf der Oberfläche richtig darstellt.	… zeigt die Evolution des Menschen so, wie sie verlaufen ist, einige Eigenschaften des Modells, z. B. die farbigen Äste, sind jedoch unnötig.	… zeigt den Wasserkreislauf insgesamt richtig, einige Details des Modells, z. B. die Farben oder die Formen, sind jedoch unnatürlich.	… stimmt mit der tatsächlichen Luftmenge in einer Lunge überein, weil das Modell auf der Basis vieler Messungen entwickelt wurde.	… stimmt mit dem Gewicht einer echten Wassermelone überein, weil das Modell auf der Basis vieler Messungen entwickelt wurde.
N II	… gleicht in einigen Merkmalen dem damals lebenden T. rex, über andere Merkmale wissen die Wissenschaftler aber nichts.	… stimmt z. B. in Bezug auf sein Kinn mit einem damals lebenden Neandertaler überein, trotzdem kann ein echter Neandertaler auch anders ausgesehen haben.	… bildet die echte Biomembran teilweise so ab, wie sie aussieht, über einige Merkmale haben die Wissenschaftler aber keine Informationen.	… sieht z. B. in Bezug auf seine Form so aus wie ein echter Grippe-Virus, trotzdem kann es in der Natur auch Grippe-Viren geben, die anders aussehen.	… stellt die Evolution des Menschen teilweise richtig dar, über weitere, noch nicht entdeckte Arten haben die Wissenschaftler aber keine Informationen.	… stellt den Wasserkreislauf richtig dar, trotzdem kann der Wasserkreislauf je nach Wetter auch anders verlaufen.	… stimmt mit der tatsächlichen Luftmenge in einer Lunge annähernd überein, das Modell trifft aber nicht auf jeden Menschen zu.	… stimmt mit dem Gewicht einer echten Wassermelone annähernd überein, Wassermelonen wachsen in der Natur aber nicht einheitlich.
	… stimmt z. B. in Bezug auf seine Körperform mit einem damals lebenden T. rex überein, trotzdem kann ein echter T. rex auch anders ausgesehen haben.	… zeigt nur wesentliche Eigenschaften des Schädelfundes, z. B. die starken Augenwülste, die breite Nase und das fliehende Kinn.	… stellt nur einen Ausschnitt der echten Biomembran dar, z. B. ihren Aufbau aus zwei Schichten oder die Anordnung der Teilchen.	… stellt nur einige Merkmale eines echten Grippe-Virus dar, z. B. die äußere Form oder die Strukturen auf der Oberfläche.	… zeigt nur einen Ausschnitt der Evolution des Menschen, weitere Schädel von früheren Vorfahren werden im Modell jedoch nicht dargestellt.	… zeigt nur wesentliche Schritte des Wasserkreislaufs, weitere Zwischenstationen des Wassers stellen die Wissenschaftler im Modell jedoch nicht dar.	... entspricht nur ungefähr der tatsächlichen Luftmenge in einer Lunge, denn das Modell bezieht nicht alle Faktoren beim Atmen in die Berechnung mit ein.	... entspricht nur ungefähr dem Gewicht einer echten Wassermelone, denn auch Umwelteinflüsse haben einen Einfluss auf deren Wachstum.
N III	... gleicht dem damals lebenden T. rex möglicherweise, sicher kann man aber nicht wissen, wie er ausgesehen hat.	... gleicht dem damals lebenden Neandertaler möglicherweise, sicher kann man aber nicht wissen, wie er ausgesehen hat.	... bildet eine Biomembran so ab, wie sie vielleicht aussieht, wie sie tatsächlich aussieht, kann man aber nicht wissen.	... bildet einen Grippe-Virus so ab, wie er vielleicht aussieht, wie er tatsächlich aussieht, kann man aber nicht wissen.	… stellt die Evolution des Menschen möglicherweise richtig dar, man kann mit dem Modell aber nur vermuten, wie sie tatsächlich verlaufen ist.	… stellt den Wasserkreislauf möglicherweise richtig dar, man kann mit dem Modell aber nur vermuten, wie er tatsächlich verläuft.	… entspricht der tatsächlichen Luftmenge in einer Lunge möglicherweise, die tatsächliche Menge an Luft kann das Modell aber nicht bestimmen.	… entspricht dem Gewicht einer echten Wassermelone möglicherweise, das tatsächliche Gewicht kann das Modell aber nicht bestimmen.
	… stimmt vielleicht mit dem damals lebenden T. rex überein, die Wissenschaftler können aber nur vermuten, wie er ausgesehen hat.	… stimmt vielleicht mit dem damals lebenden Neandertaler überein, die Wissenschaftler können aber nur vermuten, wie er ausgesehen hat.	… sieht möglicherweise so aus wie eine echte Biomembran, man kann aber nur vermuten, wie sie aussieht.	… sieht möglicherweise so aus wie ein echter Grippe-Virus, man kann aber nur vermuten, wie er aussieht.	… zeigt die Evolution des Menschen vielleicht so, wie sie verlaufen ist, sicher kann man aber nicht wissen, wie sie tatsächlich verlaufen ist.	… zeigt den Wasserkreislauf vielleicht so, wie er verläuft, sicher kann man aber nicht wissen, wie er tatsächlich verläuft.	… stimmt mit der tatsächlichen Luftmenge in einer Lunge vielleicht überein, man kann aber mit dem Modell nur vermuten, wie viel Luft in der Lunge ist.	… stimmt mit dem Gewicht einer echten Wassermelone vielleicht überein, man kann aber mit dem Modell nur vermuten, wie schwer diese ist.

Final Version of the forced choice tasks in the aspect ‘nature of models’ ($N3_3$)

Modell des *Tyrannosaurus rex* [TR]

In der linken Abbildung siehst du den Knochenfund eines T. rex (*Tyrannosaurus rex*) und in der rechten Abbildung ein Modell des T. rex, das Biologen entworfen haben.

Knochenfund eines T. rex

Modell des T. rex

Gib an, inwieweit dieses Modell des T. rex so aussieht wie ein T. rex, der vor 68 Millionen Jahren lebte!

Wähle die Aussage aus, die am ehesten deiner eigenen Meinung entspricht. Mache ein Kreuz!

Das Modell des T. rex ...

... zeigt, dass der damals lebende T. rex kurze Arme, ein großes Gebiss und einen langen Schwanz hatte.	
... stimmt z. B. in Bezug auf seine Körperform mit einem damals lebenden T. rex überein, trotzdem kann ein echter T. rex auch anders ausgesehen haben.	
... stimmt vielleicht mit dem damals lebenden T. rex überein, die Wissenschaftler können aber nur vermuten, wie er ausgesehen hat.	

Modell des Neandertalers [NT]

In der linken Abbildung siehst du den Schädel eines Neandertalers und in der rechten Abbildung ein Modell des Neandertalers, das Biologen entworfen haben.

Neandertalerschädel

Modell des Neandertalers

Gib an, inwieweit dieses Modell des Neandertalers so aussieht wie ein Neandertaler, der vor 100 000 Jahren lebte!

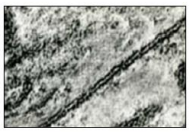
Wähle die Aussage aus, die am ehesten deiner eigenen Meinung entspricht. Mache ein Kreuz!

Das Modell des Neandertalers ...

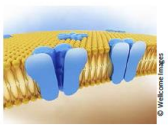
... zeigt nur wesentliche Eigenschaften des Neandertalers, z. B. die starken Augenwülste, die breite Nase und das fliehende Kinn.	
... stimmt vielleicht mit dem damals lebenden Neandertaler überein, die Wissenschaftler können aber nur vermuten, wie er ausgesehen hat.	
... zeigt, dass der Neandertaler dicke Augenwülste, eine breite Nase und ein fliehendes Kinn hatte.	

Modell der Biomembran [BM]

In der linken Abbildung siehst du eine mikroskopische Aufnahme einer Biomembran und in der rechten Abbildung ein Modell der Biomembran, das Biologen entworfen haben.



Mikroskopische Aufnahme einer Biomembran



Modell der Biomembran

Gib an, inwieweit dieses Modell der Biomembran so aussieht wie eine Biomembran, die in der Natur vorkommt!

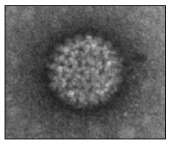
Wähle die Aussage aus, die am ehesten deiner eigenen Meinung entspricht. Mache ein Kreuz!

Das Modell der Biomembran ...

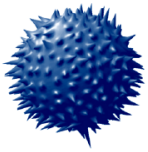
... sieht so aus wie eine vergrößerte Biomembran, weil das Modell z. B. die dicken Außenwände und den Spalt in der Mitte richtig darstellt.	
... bildet eine Biomembran so ab, wie sie vielleicht aussieht, wie sie tatsächlich aussieht, kann man aber nicht wissen.	
... bildet die echte Biomembran teilweise so ab, wie sie aussieht, über einige Merkmale haben die Wissenschaftler aber keine Informationen.	

Modell des Grippe-Virus [VS]

In der linken Abbildung siehst du eine mikroskopische Aufnahme eines Grippe-Virus und in der rechten Abbildung ein Modell des Grippe-Virus, das Biologen entworfen haben.



Mikroskopische Aufnahme eines Grippe-Virus



Modell des Grippe-Virus

Gib an, inwieweit dieses Modell des Grippe-Virus so aussieht wie ein Grippe-Virus, der in der Natur vorkommt!

Wähle die Aussage aus, die am ehesten deiner eigenen Meinung entspricht. Mache ein Kreuz!

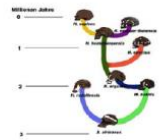
Das Modell des Grippe-Virus ...

... bildet einen Grippe-Virus so ab, wie er vielleicht aussieht, wie er tatsächlich aussieht, kann man aber nicht wissen.	
... sieht z. B. in Bezug auf seine Form so aus wie ein echter Grippe-Virus, trotzdem kann es in der Natur auch Grippe-Viren geben, die anders aussehen.	
... bildet einen echten Grippe-Virus so ab, wie er aussieht, weil es von Wissenschaftlern mit Hilfe der neuesten Technik hergestellt wurde.	

Modell des menschlichen Stammbaums [EV]

In der linken Abbildung siehst du verschieden alte Schädel von menschlichen Vorfahren und in der rechten Abbildung ein Modell des menschlichen Stammbaums, das Biologen entworfen haben.

Schädel von menschlichen Vorfahren



Modell des menschlichen Stammbaums

Gib an, inwieweit das Modell des menschlichen Stammbaums die menschliche Evolution zutreffend darstellt!

Wähle die Aussage aus, die am ehesten deiner eigenen Meinung entspricht. Mache ein Kreuz!

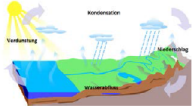
Das Modell des menschlichen Stammbaums ...

... zeigt nur einen Ausschnitt der Evolution des Menschen, weitere Schädel von früheren Vorfahren werden im Modell jedoch nicht dargestellt.	
... zeigt die Evolution des Menschen so, wie sie verlaufen ist, einige Eigenschaften des Modells, z. B. die farbigen Äste, sind jedoch unnötig.	
... zeigt die Evolution des Menschen vielleicht so, wie sie verlaufen ist, sicher kann man aber nicht wissen, wie sie tatsächlich verlaufen ist.	

Modell des Wasserkreislaufs [WK]

In der linken Abbildung siehst du ein Foto eines Regenschauers und in der rechten Abbildung ein Modell des Wasserkreislaufs, das Biologen entworfen haben.

Foto eines Regenschauers



Modell des Wasserkreislaufs

Gib an, inwieweit das Modell des Wasserkreislaufs den in der Natur ablaufenden Kreislauf des Wassers zutreffend darstellt!

Wähle die Aussage aus, die am ehesten deiner eigenen Meinung entspricht. Mache ein Kreuz!

Das Modell des Wasserkreislaufs ...

... zeigt den Wasserkreislauf vielleicht so, wie er verläuft, sicher kann man aber nicht wissen, wie er tatsächlich verläuft.	
... zeigt den Wasserkreislauf insgesamt richtig, einige Details des Modells, z. B. die Farben oder die Formen, sind jedoch unnatürlich.	
... zeigt nur wesentliche Schritte des Wasserkreislaufs, weitere Zwischenstationen des Wassers stellen die Wissenschaftler im Modell jedoch nicht dar.	

Version 1 of the answer options in the aspect ‘purpose of models’ ($P1_6$)

Modell	T. rex	Jurawald	Biomembran	Gehirn	Evolution	Uferzone	Luftstrom	Bakt.wachstum
N I	… die Fortbewegungsart des T. rex anschaulich darzustellen.	… den grundsätzlichen Aufbau des Jurawaldes darzustellen.	… den Aufbau der Biomembran vergrößert darzustellen.	… die Lage der verschiedenen Gehirnareale darzustellen.	… die menschliche Evolution vereinfacht abzubilden.	… Vertreter typischer Pflanzenarten an einer Uferzone zusammen zu stellen.	… die Luftmenge in der Lunge zu einem bestimmten Zeitpunkt wiederzugeben.	… die Menge an Bakterien auf einem Nährmedium wiederzugeben.
	… den Körperbau des ausgestorbenen T. rex möglichst genau zu zeigen.	… die verschiedenen Stockwerke und Pflanzenarten des Jurawaldes zu zeigen.	… die verschiedenen Bestandteile der Biomembran zu veranschaulichen.	… die Funktionen der einzelnen Gehirnareale zu veranschaulichen.	… die Schädel der wichtigsten menschlichen Vorfahren zusammen zu stellen.	… die verschiedenen Bereiche der Uferzone abzubilden.	… die Menge an Luft in der Lunge beim Ein- und Ausatmen zu beschreiben.	… das Wachstum von Bakterien auf einem Nährmedium zu beschreiben.
N II	… den Zusammenhang zwischen Knochen und Körperbau zu erläutern.	… das Zusammenleben der Pflanzen im Jurawald verständlich zu machen.	… den Aufbau der Biomembran zu erläutern.	… den Zusammenhang von Lage und Funktion der Gehirnareale zu erläutern.	… die Abstammungsverhältnisse zwischen den Schädeln zu verdeutlichen.	… die Abhängigkeit bestimmter Pflanzenarten von Wasser zu verdeutlichen.	… die Luftmenge in der Lunge beim Ein- und Ausatmen zu berechnen.	… den Zusammenhang zwischen Wachstumsfaktor und Anzahl an Bakterien auszuführen.
	… den Einfluss der Vorderbeine auf die Fortbewegung verständlich zu machen.	… die Entwicklung verschiedener Pflanzenarten im Jurawald zu erläutern.	… das Zusammenwirken der Bestandteile begreiflich zu machen.	… die Funktionsweise der verschiedenen Gehirnareale begreiflich zu machen.	… die Evolution des Menschen nachvollziehbar zu machen.	… das Zusammenleben von Pflanzenarten in der Uferzone nachvollziehbar zu machen.	… den Zusammenhang zwischen Stärke und Häufigkeit der Atmung auszuführen.	… die Anzahl an Bakterien auf einem Nährmedium zu berechnen.
N III	… den Körperbau des ausgestorbenen T. rex zu untersuchen.	… das Wachstum einiger Pflanzen im Jurawald zu untersuchen.	… den Aufbau der Biomembran zu untersuchen.	… den Aufbau des Gehirns zu untersuchen.	… die Verwandtschaftsverhältnisse menschlicher Vorfahren aufzuklären.	… die pflanzliche Besiedelung einer Uferzone aufzuklären.	… Vermutungen über die Luftmenge in der Lunge zu überprüfen..	… Vermutungen über die Anzahl an Bakterien auf einem Nährmedium zu überprüfen.
	… die Fortbewegungsart des T. rex weiter zu erkunden.	… die Beziehungen zwischen verschiedenen Pflanzen im Jurawald zu erkunden.	… die einzelnen Bestandteile der Biomembran zu erkunden.	… die Funktionsweise der einzelnen Gehirnareale aufzudecken.	… die weitere Evolution des Menschen vorauszusagen.	… den Standort einer bestimmten Pflanzenart vorauszusagen.	… Vorhersagen über die Menge an Luft in der Lunge abzuleiten.	… Vorhersagen über die Anzahl an Bakterien auf einem Nährmedium abzuleiten.

Version 2 of the answer options in the aspect ‘purpose of models ($P2_6$)

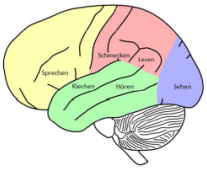
Modell	T. rex	Jurawald	Biomembran	Gehirn	Evolution	Uferzone	Luftstrom	Bakt.wachstum
N I	… die Fortbewegungsart des ausgestorbenen T. rex anschaulich darzustellen.	… den grundsätzlichen Aufbau des Jurawaldes darzustellen.	… den Aufbau der Biomembran sichtbar zu machen.	… die Lage der verschiedenen Gehirnareale sichtbar zu machen.	… die Schädel der wichtigsten menschlichen Vorfahren zusammen zu stellen.	… Vertreter typischer Pflanzenarten an einer Uferzone zusammen zu stellen.	… die Luftmenge in der Lunge wiederzugeben.	… die Menge an Bakterien auf einem Nährmedium wiederzugeben.
	… den Körperbau des T. rex möglichst genau zu zeigen.	… die verschiedenen Stockwerke und Pflanzenarten des Jurawaldes zu zeigen.	… die verschiedenen Bestandteile der Biomembran zu veranschaulichen.	… die Funktionen der einzelnen Gehirnareale zu veranschaulichen.	… die menschlichen Abstammungs-verhältnisse vereinfacht abzubilden.	… die verschiedenen Bereiche der Uferzone abzubilden.	… die Menge an Luft in der Lunge zu beschreiben.	… die Anzahl an Bakterien auf einem Nährmedium zu beschreiben.
N II	… den Zusammenhang zwischen Knochen und Körperbau des T. rex zu erläutern.	… die Entwicklung verschiedener Pflanzenarten im Jurawald zu erläutern.	… den Aufbau der Biomembran zu erklären.	… den Zusammenhang von Lage und Funktion der Gehirnareale zu erklären.	… die Abstammungs-verhältnisse menschlicher Vorfahren zu erklären.	… die Abhängigkeit verschiedener Pflanzenarten von Wasser zu erklären.	… die Luftmenge in der Lunge beim Ein- und Ausatmen zu bestimmen.	… das Wachstum von Bakterien auf einem Nährmedium zu bestimmen.
	… den Einfluss der Vorderbeine auf die Fortbewegungsart des T. rex verständlich zu machen.	… das Zusammenleben der Pflanzen im Jurawald verständlich zu machen.	… das Zusammenwirken der Bestandteile begreiflich zu machen.	… die Funktionen der verschiedenen Gehirnareale begreiflich zu machen.	… Beziehungen zwischen den verschiedenen Schädeln nachvollziehbar zu machen.	… das Zusammenleben von Pflanzenarten in der Uferzone nachvollziehbar zu machen.	… den Zusammenhang zwischen Stärke und Häufigkeit der Atmung auszuführen.	… den Zusammenhang zwischen Wachstum und Anzahl an Bakterien auszuführen.
N III	… den Körperbau des ausgestorbenen T. rex zu erforschen.	… die Beziehungen zwischen verschiedenen Pflanzenarten im Jurawald zu erforschen.	… den Aufbau der Biomembran weiter zu erforschen.	… Lage der verschiedenen Gehirnareale weiter zu erforschen.	… die Abstammungs-verhältnisse menschlicher Vorfahren aufzuklären.	… die pflanzliche Besiedelung einer Uferzone aufzuklären.	… Vermutungen über die Luftmenge in der Lunge aufzustellen.	… Vermutungen über die Menge an Bakterien auf einem Nährmedium aufzustellen.
	… Vermutungen über die Fortbewegungsart des T. rex abzuleiten.	… Vermutungen über das Wachstum einiger Pflanzen im Jurawald abzuleiten.	… weitere Bestandteile der Biomembran vorauszusagen.	… die Funktionen der einzelnen Gehirnareale vorauszusagen.	… die weitere Entwicklung des menschlichen Schädels vorherzusagen.	… den Standort einer bestimmten Pflanzenart vorherzusagen.	… Vorhersagen über die Menge an Luft in der Lunge abzuleiten.	… Vorhersagen über die Anzahl an Bakterien auf einem Nährmedium abzuleiten.

Final Version of the forced choice tasks in the aspect ‘purpose of models’ ($P3_3$)

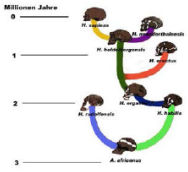
Modell des *Tyrannosaurus rex* [TR]

In der Abbildung siehst du ein Modell des T. rex, das Biologen entworfen haben.	
 Modell des T. rex	
Modelle werden für einen bestimmten Zweck entwickelt. Gib an, welchen Zweck dieses Modell des T. rex haben kann!	
Wähle die Aussage aus, die am ehesten deiner eigenen Meinung entspricht. Mache ein Kreuz!	
Das Modell des T. rex hat den Zweck ...	
… den Körperbau des T. rex möglichst genau zu zeigen.	
… Vermutungen über die Fortbewegungsart des T. rex abzuleiten.	
… den Einfluss der Vorderbeine auf die Fortbewegungsart des T. rex verständlich zu machen.	

Modell des Gehirns [GH]

In der Abbildung siehst du ein Modell des menschlichen Gehirns, das Biologen entworfen haben.	
 Modell des menschlichen Gehirns	
Modelle werden für einen bestimmten Zweck entwickelt. Gib an, welchen Zweck dieses Gehirnmodell haben kann!	
Wähle die Aussage aus, die am ehesten deiner eigenen Meinung entspricht. Mache ein Kreuz!	
Das Modell des Gehirns hat den Zweck ...	
… die Lage der verschiedenen Gehirnareale weiter zu erforschen.	
… die Lage der verschiedenen Gehirnareale sichtbar zu machen.	
… den Zusammenhang von Lage und Funktion der Gehirnareale zu erklären.	

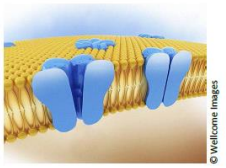
Modell des menschlichen Stammbaums [EV]

In der Abbildung siehst du ein Modell des menschlichen Stammbaums, das Biologen entworfen haben.	
 Modell des menschlichen Stammbaums	
Modelle werden für einen bestimmten Zweck entwickelt. Gib an, welchen Zweck dieses Stammbaummodell haben kann!	
Wähle die Aussage aus, die am ehesten deiner eigenen Meinung entspricht. Mache ein Kreuz!	
Das Modell des menschlichen Stammbaums hat den Zweck ...	
… die menschlichen Abstammungsverhältnisse vereinfacht abzubilden.	
… die Abstammungsverhältnisse menschlicher Vorfahren zu erklären.	
… die weitere Entwicklung des menschlichen Schädels vorherzusagen.	

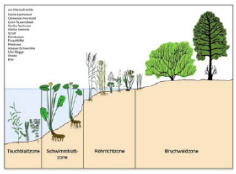
Modell des Jurawaldes [JW]

In der Abbildung siehst du ein Modell des Waldes aus dem Jurazeitalter, das Biologen entworfen haben.	
 Modell des Jurawaldes	
Modelle werden für einen bestimmten Zweck entwickelt. Gib an, welchen Zweck dieses Modell des Jurawaldes haben kann!	
Wähle die Aussage aus, die am ehesten deiner eigenen Meinung entspricht. Mache ein Kreuz!	
Das Modell des Jurawaldes hat den Zweck ...	
… Vermutungen über das Wachstum einiger Pflanzen im Jurawald abzuleiten.	
… die Entwicklung verschiedener Pflanzenarten im Jurawald zu erläutern.	
… den grundsätzlichen Aufbau des Jurawaldes darzustellen.	

Modell der Biomembran [BM]

In der Abbildung siehst du ein Modell der Biomembran, das Biologen entworfen haben.	
 Modell der Biomembran	
Modelle werden für einen bestimmten Zweck entwickelt. Gib an, welchen Zweck dieses Modell der Biomembran haben kann!	
Wähle die Aussage aus, die am ehesten deiner eigenen Meinung entspricht. Mache ein Kreuz!	
Das Modell der Biomembran hat den Zweck ...	
… den Aufbau der Biomembran zu erklären.	
… den Aufbau der Biomembran sichtbar zu machen.	
… den Aufbau der Biomembran weiter zu erforschen.	

Modell der Uferzone [UF]

In der Abbildung siehst du ein Modell der Uferzone um einen See, das Biologen entworfen haben.	
 Modell der Uferzone	
Modelle werden für einen bestimmten Zweck entwickelt. Gib an, welchen Zweck dieses Modell der Uferzone haben kann!	
Wähle die Aussage aus, die am ehesten deiner eigenen Meinung entspricht. Mache ein Kreuz!	
Das Modell der Uferzone hat den Zweck ...	
… die Abhängigkeit verschiedener Pflanzenarten von Wasser zu erklären.	
… den Standort einer bestimmten Pflanzenart vorherzusagen.	
… die verschiedenen Bereiche der Uferzone abzubilden.	

9.6. Attachments

Attachment 1. Overview of studies assessing students' meta-modelling knowledge with closed-ended instruments. Information concern sample (left column), assessment instrument (middle column) and findings (right column).

Grosslight, Jay, Unger & Smith. (1991). Understanding models and their use in science: Conceptions of middle and high school students and experts.		
33 students from grade 7, 22 students from grade 11 and 4 experts	Series of questions about models: What comes to mind when you hear the word 'model? Are there different kinds of models? What are models for? Can you use models in science? What do you have to think about when making a model? Do you think scientists would ever have more than one model for the same thing? Would a scientist ever change a model? (Four physical items: toy airplane, subway map, picture of a house, schematic diagram of the water cycle).	Aspects: kinds of models, multiple models for the same thing, purpose of models, designing and creating models, and changing a model. Global levels of students' meta-modelling knowledge: (I) models are simple copies of reality with the purpose of matching the real thing; (II) construction of model is linked to a specific purpose and cannot be exact duplicates of reality; (III) models are tools for testing and revising hypotheses about the corresponding phenomenon. Grade 7: Level I: 67%, Level II: 12%, Level III: 0%. Grade 11: Level I: 23 %, Level II: 36 %, Level III: 0%.
Harrison & Treagust (1996). Secondary students' mental models of atoms and molecules: Implications for teaching chemistry.		
48 students in grades 8 to 10 (plus some grade 11 students)	Semi structured focused interviews about analogical models of molecules used in chemistry instruction. Students from 11th grade are being reported from an unpublished study in 1995.	Each student was assigned as a level I, II, or III modeler according to the levels described by Grosslight et al., (1991). Level I: 58 %; Level II: 42 %; Level III: 0 %.
Treagust, Chittleborough & Mamiala (2002). Students' understanding of the role of scientific models in learning science.		
228 students from grade 8 to grade 10	Students' Understanding of Models in Science (SUMS). 27 decontextualised items to be rated on a five-point Likert-type scale. 5 factors: 'Models as multiple representtations'; 'Models as exact replicas'; 'Models as explanatory tools'; 'Uses of scientific models'; and 'The changing nature of models'.	Students (> 50 %) recognised need for multiple models for particular aspects of the item as well as for individual needs of the learner. Students (75 %) agree that a model needs to be close to the real thing. Students (> 65 %) primarily value this descriptive aspect of models. Half of students agreed that scientific models are used for making predictions, formulating theories and showing how information is used. Students (> 70 %) state models will change according to changes in scientific thinking. There were no statistically significant differences across grades.

Treagust, Chittleborough & Mamiala (2004). Students' Understanding of the Descriptive and Predictive Nature of Teaching Models in Organic Chemistry.

36 students from grade 11	Molecular Representations (MR) questionnaire: Students respond on a 5-point Likert scale how much they agree with 11 purposes attributed to each of 4 different teaching models. My Views of Models and Modelling in Science (VOMMS) questionnaire: Students choose between two alternative statements about scientific models and then explain their choice.	MR: Majority of students see purpose in accurate depiction of attributes, including limitations and strengths. Few students (< 3 %) indicated that models can be used to make and test predictions. VOMMS: Students (> 80 %) see models as representations of ideas of how things work; accept multiple models to explain ideas; state that models are used to explain phenomena; that a model is based on the facts that support the theory; that a model is accepted when it can be used to explain results and that a model may change in future years.

Chittleborough, Treagust, Mamiala & Mocerino (2005). Students' perceptions of the role of models in the process of science and in the process of learning.

275 students from grade (G) 8 to first-year of university (U)	My Views of Models and Modelling in Science (VOMMS).	Significant differences across the grades describing a model as an 'accurate duplicate of reality' (G8: 25 %, G11: 8 %; U: 12 %). Significant differences between the grades agreeing that 'one model only' is preferable to explain scientific ideas (G8: 31 %, G11: 3 %, U: 0 %). Students (> 70 %) agreed that 'a model is accepted on the facts that support it and the theory', 'when it can explain results' and that 'scientific models will change in the future'.

Sins, Savelsbergh, van Joolingen & van Hout-Wolters (2009). The Relation between Students' Epistemological Understanding of Computer Models and their Cognitive Processing on a Modelling Task.

26 students from grade 11	Open-ended questionnaire with 10 tasks to measure students' epistemological understanding of the nature of models, the purpose of models, the design and revision of models, and the evaluation of models.	Each student was given a level of epistemological understanding rating (high, moderate, low) per dimension according to the three levels by Grosslight et al. (1991). 'Nature of models' (Level I: 23 %; Level II: 65 %; Level III: 12 %); 'Purpose of models' (Level I: 12 %; Level II: 46 %; Level III: 42 %).

Al-Balushi (2011). Students' evaluation of the credibility of scientific models that represent natural entities and phenomena.

845 students in grades (G) 9 to 11 and 108 preservice science teachers at years 3–5	Epistemologies about the Credibility of Scientific Models (ECMS) based on a Certainty, Imaginary, Suspicion, Denial taxonomy. Students rate a list of natural entities and phenomena that are located at different points along the concrete–abstract continuum.	The overall students' responses to the ECSM survey across grade levels showed a decrease in the certainty level (G9: 44 % to G11: 29 %) and an increase in the imaginary level (G9: 37 % to G11: 46 %). Competing concrete–abstract parallel nature of some scientific models may be responsible for non-typical epistemological perceptions.

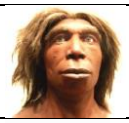
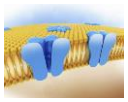
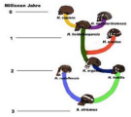
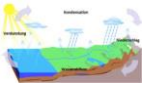
Krell (2013). Wie Schülerinnen und Schüler biologische Modelle verstehen.

1216 students in grades 7 to 10	Contextualised forced choice tasks for each of the five aspects of the 'model of model competence'. Students asked to bring three statements (each representing one of the three levels of the theoretical framework) into a preference rank order.	Majority of students understands models as idealised representations (Level II: 52 %). Students believed an original to allows the creation of different models (Level II: 50 %) and a model to serve the purpose of predicting something about the original (Level III: 37 %). Students recognised the possibility to test hypotheses about the original with the model (Level III: 46 %) and believed, a model should be changed due to new findings about the original (Level II: 50 %).

Grünkorn (2014). Modellkompetenz im Biologieunterricht: Empirische Analyse von Modellkompetenz bei Schülerinnen und Schülern der Sekundarstufe I mit Aufgaben im offenen Antwortformat.

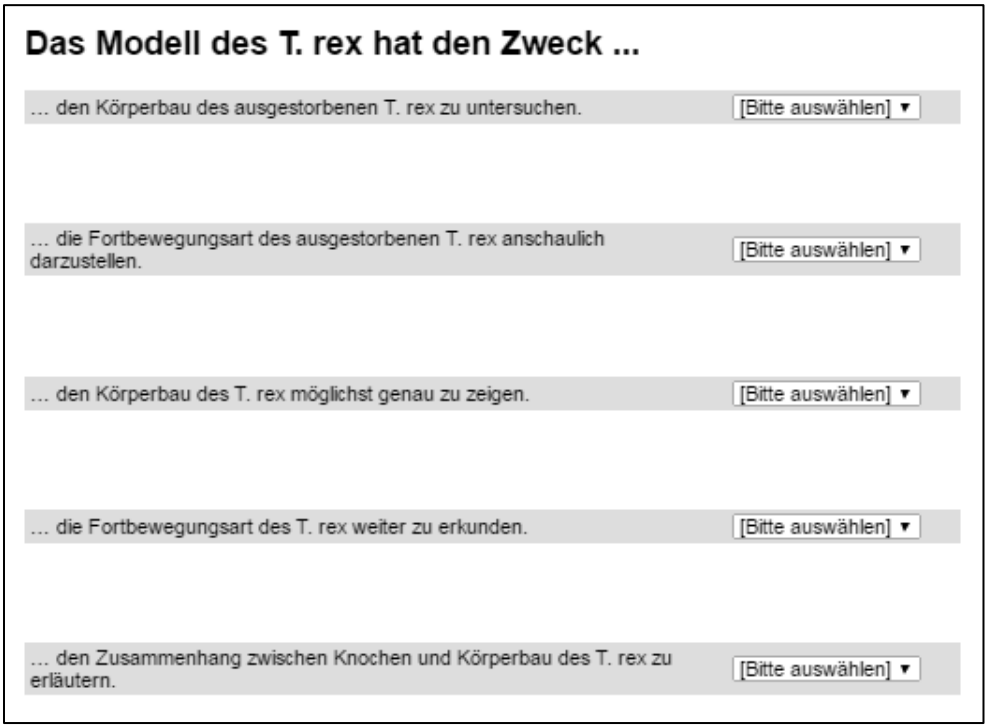
1177 students in grades 7 to 10	Contextualised open-ended tasks for each of the five aspects of the 'model of model competence'.	Majority of students understands models as replications of an original (Level I: 75 %), which have different model object properties (Level I: 44 %), which serve for showing facts (Level I: 52 %), which, in order to be tested, should be compared to an original (Level II: 67 %) and which should be changed when they do not match the original (Level II: 61 %).

Attachment 2. Models which have been used in the forced choice tasks. Those models used in the final diagnostic instrument are highlighted in grey. N = 'nature of models'; P = 'purpose of models'.

Model	Picture	Typology	Aspect	Licence
TR (*Tyrannosaurus rex*)		concrete	N / P	Ryanz720. Trexg. Public domain. https://commons.wikimedia.org/wiki/File:Trexg.jpg
NT (Neanderthal man)		concrete	N	Juliane Grünkorn. Neanderthal man in the natural historical museum Berlin. Rights obtained.
AR (Archaeopteryx)		concrete	N	http://foter.com/photo/archaeopteryx-2b/
JF (Jura forest)		concrete	P	Sarah Gogolin. Fern in the botanical garden Berlin. Own picture.
BM (bio membrane)		concrete-abstract	N / P	Maurizio De Angelis. Ion channels. CC BY-NC-ND 2.0. https://www.flickr.com/photos/wellcomeimages/5814248573
VS (influenza virus)		concrete-abstract	N	Hitthatswitch. Swine Flu H5N1 virus influenza. CC BY-NC-SA 2.0. https://www.flickr.com/photos/ringai/3912577366
BR (brain)		concrete-abstract	P	NEUROtiker. Seitenansicht eines menschlichen Gehirns. CC-BY-SA-2.5. https://commons.wikimedia.org /wiki/File:Gehirn,_lateral_Lobi_deu.svg
EV (evolution)		abstract-concrete	N / P	Tierdoku. Stammbaum-mensch-2054. CC BY-SA 3.0. http://tierdoku.com/index.php?title=Bild:Stammbaum-mensch-2054.jpg
WC (water circle)		abstract-concrete	N	Joooo. Wasserkreislauf. CC BY-SA 3.0. http://commons.wikimedia.org/wiki/File:Water_Cycle_-_blank.svg
LZ (lakeshore zone)		abstract-concrete	P	No licence. http://www.uni-duesseldorf.de/MathNat/Biologie/Didaktik/WasserSek_I/oekosystem_see/bilder/uferzonen_kl.jpg
AS (air stream)	$f(x) = \mathrm{k}\sin(ax)$	abstract	N / P	Sarah Gogolin. Sinus curve. Own drawing.
WM (water melon)	$f(x) = c \cdot a^x$	abstract	N	Sarah Gogolin. Formula. Own drawing.
BG (bacterial growth)	$N_t = N_0 * e^{\lambda t}$	abstract	P	Sarah Gogolin. Formula. Own drawing.

Note: The typology refers to Krell et al.'s (2014b) student-based typology of biological models.

Attachment 3. Screenshot of the answer options in the expert judgement questionnaire for the aspect ‘purpose of models’. The screenshot is cut after five of seven answer options (one answer option was displayed twice so to prevent guessing the last given level).



Attachment 4. Screenshot of the answer bracket for the additional questions in the expert judgement questionnaire for the aspect ‘purpose of models’.

Attachment 5. Instructions for the interviewer for concurrent thinking aloud and retrospective descriptions of thinking (‘nature of models’; study IV).

Leitfaden für lautes Denken und retrospektive Interviews

FC-Aufgaben Einführung

Ich bin [...] und ich schreibe meine Doktorarbeit an der Freien Universität zu Modellen in der Biologie. Dabei interessiert mich vor allem, was Schüler so über Modelle denken.
Um das herauszufinden, habe ich Aufgaben zu verschiedenen Modellen entwickelt. Einige dieser Aufgaben werde ich dir gleich stellen.
Grundsätzlich interessiert mich nur, was du persönlich denkst. Richtige oder falsche Aussagen gibt es deswegen nicht. Du kannst also keine Fehler machen oder irgendwas Falsches sagen.
Damit ich genau weiß, was du bei der Bearbeitung der Aufgaben denkst, würde ich gerne eine Methode benutzen, die sich *lautes Denken* nennt.

Lautes Denken Einführung und Übung

Beim *lauten Denken* geht es darum, genau herauszufinden, was du beim Bearbeiten einer Aufgabe denkst.
Deswegen würde ich dich bitten, alles zu erzählen, was dir durch den Kopf geht, wenn du die Aufgabe bearbeitest.
Zu diesem Zweck, lies bitte auch alles laut vor.
Im Idealfall sprichst du die ganze Zeit ohne Unterbrechung, also möglichst pausenlos. Ok?

Um das Ganze mal zu üben, werde ich dir jetzt eine Frage stellen und dich bitten, laut zu denken.
Versuche zu schätzen, wie viele Türen in deiner Wohnung sind und erzähle alles, was dir dabei durch den Kopf geht. [...]
Vielen Dank. Hast du das Prinzip verstanden?

Aufgabe 1 A/B beantworten

Lautes Denken

Ich lege dir jetzt zwei Aufgaben zum Modell [...] vor und würde dich bitten, sie zu beantworten.
Vergiss dabei nicht, alles laut zu äußern, was dir durch den Kopf geht.
Bist du bereit?

Modell eines *Tyrannosaurus rex*

In der linken Abbildung siehst du den Knochenfund eines T. rex (*Tyrannosaurus rex*) und in der rechten Abbildung ein Modell eines T. rex, das Biologen entworfen haben.

Abbildung 1: Knochenfund eines T. rex

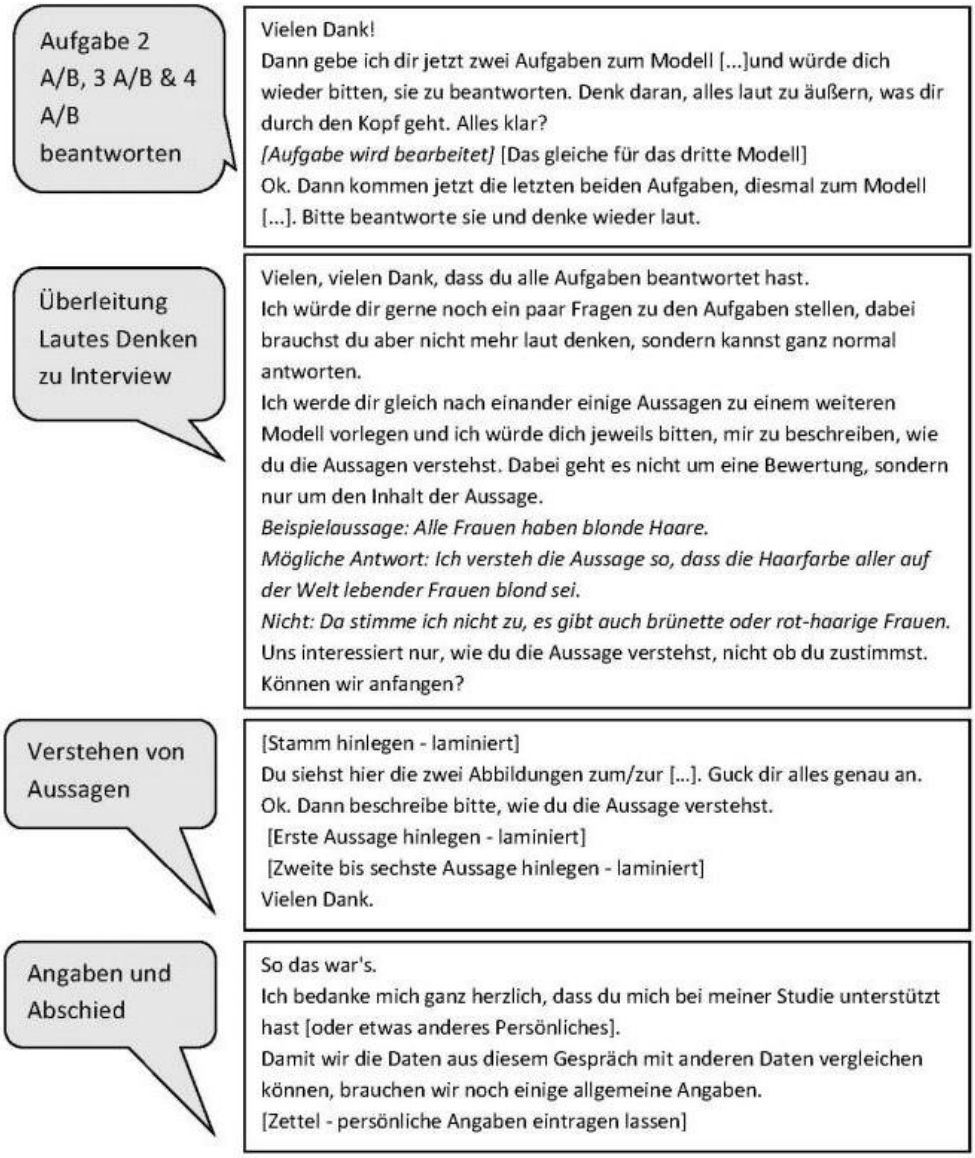
Abbildung 2: Modell eines T. rex

Gib an, inwieweit dieses Modell des T. rex so aussieht wie ein T. rex, der vor 68 Millionen Jahren lebte!

Wähle die Aussage aus, die am ehesten deiner eigenen Meinung entspricht. Kreuze an!	
Das Modell des T. rex ...	
... stimmt mit dem damals lebenden T. rex überein, weil das Modell eine Nachbildung des Knochenfundes ist.	
... gleicht in einigen Merkmalen dem damals lebenden T. rex, über andere Merkmale wissen die Wissenschaftler aber nichts.	
... gleicht dem damals lebenden T. rex möglicherweise, sicher kann man aber nicht wissen, wie er ausgesehen hat.	

Wähle die Aussage aus, die am ehesten deiner eigenen Meinung entspricht. Kreuze an!	
Das Modell des T. rex ...	
... stimmt vielleicht mit dem damals lebenden T. rex überein, die Wissenschaftler können aber nur vermuten, wie er ausgesehen hat.	
... zeigt, dass der damals lebende T. rex kurze Arme, ein großes Gebiss und einen langen Schwanz hatte.	
... stimmt z. B. in Bezug auf seine Körperform mit einem damals lebenden T. rex überein, trotzdem kann ein echter T. rex auch anders ausgesehen haben.	

[Continuation from previous page]



Verteilung der Modelle auf Proband_innen

	TR	NT	EV	WK	BM	VS	LS	WM
1	x	x		x		x	**X**	
2		x	x		x		x	**X**
3	**X**		x	x		x		x
4	x	**X**		x	x		x	
5		x	**X**		x	x		x
6	x		x	**X**		x	x	
7		x		x	**X**		x	x
8	x		x		x	**X**		x

x=Lautes Denken; **X**=Retrospektives Interview

Attachment 6. Coding scheme (0/1) for the raters of concurrent think aloud protocols and retrospective descriptions of thinking (‘nature of models’; study IV).

(1) Es werden nur Aussagenteile beachtet, die sich auf das Verhältnis zwischen Original und Model beziehen.

- Niveau I: Ein Modell ist (nahezu) identisch, sehr ähnlich, besitzt zum Original nur Unterschiede im Maßstab, …
- Niveau II: Ein Modell ist idealisiert, vereinfacht, zeigt eingeschränkte, teilweise Gleichheit, zeigt Gleichheit nur in Bezug auf wenige/wesentliche Aspekte/Schwerpunkte, …
- Niveau III: Ein Modell ist: Gedankenmodell, Hypothese, Theorie, Rekonstruktion, …

	Niveau I	**Niveau II**	**Niveau III**
Eigenschaften von Modellen	Modelle sind Kopien *von* etwas	Modelle sind idealisierte Repräsentationen *von* etwas	Modelle sind theoretische Rekonstruktionen *von* etwas

(2) Die Schülerantworten werden den Kodes (0/1) zugeordnet. 0 = nicht auf dem intendierten Niveau verstanden; 1 = auf dem intendierten Niveau verstanden.

(3) Sollte eine Aussage keine Informationen darüber beinhalten, ob der/ die Schüler*in die Antwortalternative auf dem entsprechenden theoretischen Niveaus versteht und interpretiert wird mit (3) kodiert, was der Kategorie „Antwort nicht aussagekräftig“ gleichkommt.

(4) Pro Schülerantwort wird nur ein Kode zugeordnet.

(5) Bei der Kodierung wird jede Schülerantwort separat betrachtet und Bezüge zu einer vorherigen Aussage werden außer Acht gelassen.

(6) Sollte ein/e Schüler*in die Antwortalternative nur wiederholen oder nur Schlüsselbegriffe aus der Antwortalternative ohne weiter gehende Erläuterung nennen, wird mit (3) kodiert.

Attachment 7. Coding scheme (0/1) for the raters of retrospective interviews ('purpose of models'; study VI). [Here, in order to not be repetative, I only display those parts of the coding scheme that differ from the one presented in Attachment 6].

(1) Es werden nur Aussagenteile beachtet, die sich auf den Zweck des Models beziehen.

Niveau	Erläuterung	Ankerbeispiel [von Grünkorn]
I	**Modell zum Darstellen eines Sachverhalts** stellt Sachverhalte dar	• *Das Modell zeigt die verschiedenen Pflanzen, die [in] einem Wald vorkommen.*
II	**Modell zum Erkennen und Erklären von Zusammenhängen** beschreibt und erklärt Zusammenhänge zwischen verschiedenen Aspekten im Original und dient dazu, bekannte Tatsachen nachzuvollziehen	• *Das Modell stellt dar, dass sie gucken kann, wie die Blätter und Blüten sich entwickeln und sich verbreiten.* • *Den Aufbau eines Waldes kann man daran auch erklären. Pflanzen (also auch Wälder) brauchen Erde, damit sie ihre Wurzeln ausschlagen können.*
III	**Modell zum Überprüfen von Ideen** dient als Instrument zur Überprüfung von Hypothesen über das Original dient dazu, Schlüsse über das Original zu ziehen dient dazu, Erkenntnisse über das Original auf andere Phänomene zu übertragen	• *Vielleicht kann man gewisse Tests durchführen und so die Wirkung von bestimmten Dingen überprüfen.* • *So ein Modell ist wahrscheinlich dazu da, um zu überprüfen, ob sich Pflanzen „vermehren" können. Dies braucht man wiederum, um etwas über den Wald zu sagen und den Bestand dann in Zahlen oder in Diagrammen festzuhalten. Vielleicht kann man das dann auf andere Ökosysteme wie das Meer oder so beziehen.*

Das Modell der Biomembran hat den Zweck, weitere Bestandteile der Biomembran vorauszusagen. [3.2]

Kodierung	Ankerbeispiel
1 - Antwortalternative wurde entsprechend der theoretischen Intention verstanden	**bean2201:** Ich denke, das Modell dient nur dazu, das Ganze zu veranschaulichen, aber man kann anhand eines Modells keine Vorhersagen treffen. **maon0701:** Weil ich dachte, da geht's darum, dass man jetzt sagen kann, in einem bestimmten Abstand kommt jetzt bestimmt wieder so ein blaues Dings-Bums da.
0 - Antwortalternative wurde nicht entsprechend der theoretischen Intention verstanden	**kaix2401:** Man erkennt im Modell zwar weitere, die sind aber nicht beschriftet oder irgendwas. Und vorauszusagen war auch so ein bisschen komisch formuliert. - *Was findest du daran komisch?* - Naja vorauszusagen. Wenn, dann hätte ich jetzt anzugeben oder sowas eingesetzt, weil das ist irgendwie ein bisschen klarer dann.

(2) Niveauspezifische Regeln für die Kodierung der Schülerantworten

- Der Proband hat die Aussage auf dem intendierten Niveau verstanden.

Niveau	Beispiel Antwort-alternative	Schüler*innenaussage	Begründung/Regel
I	...die verschiedenen Bestandteile der Biomembran zu veranschaulichen.	faia3107: Meiner Meinung nach wird das da ziemlich gut dargestellt.	„Darstellen"/ „zeigen"/ „sichtbar machen" / „repräsentieren" / „beschreiben" / „ausführen" / „abbilden" / „wiedergeben" / „nachahmen" / „darlegen" / „charakterisieren"/ … sind im Rahmen der theoretischen Grundlage zum Zweck von Modellen ein Synonym zu „veranschaulichen".

	… die Lage der verschiedenen Gehirnareale sichtbar zu machen.	anas0311: Da in diesem Modell unter eine Farbe verschiedene Textgruppen kommen.	Durch visuelle Aspekte, wie in diesem Fall „Farbe“ werden in dem Modell verschiedene Bereiche sichtbar gemacht. Ebenfalls werden hier als verstanden Antworten gewertet, die sich deskriptiv auf den äußeren Aufbau beziehen.
II	… die Abstammungsverhältnisse menschlicher Vorfahren zu erklären.	hill0109: Da wir hier die Entwicklung des Menschen nachvollziehen können. bean2201: weil in so einem Modell ist viel mehr und viel leichter zu erkennen als auf so einem Foto.	Das Modell wird als Mittel anerkannt, mit dem man Wissen begreiflich machen kann. Aussagen, die erklären mit bspw. leichter/ besser/ mehr… zu erkennen/ sehen beinhalten, werden als Niveau II gewertet, da das beschreibende Adjektiv in Verbindung mit „erkennen“ über Niveau I hinausgeht.
	...den Zusammenhang zwischen Stärke und Häufigkeit der Atmung auszuführen.	clis0403: wenn die Stärke der Atmung erhöht wird und die Häufigkeit herabgesetzt wird, könnte sich das auf die Luftmenge in der Lunge auswirken.	Der Inhalt der Antwortalternative wird exemplarisch erklärt. Dementsprechend wurde erkannt, was der Inhalt der Antwortalternative ist.
III	...Vorhersagen über die Anzahl an Bakterien auf einem Nährmedium abzuleiten.	anli1609: Also irgendwie vermutet man ja nicht oder leitet irgendwie ab. Oder vorhersagen halt.	„Ableiten“ wurde im richtigen Sinn des Wortes – als etwas Unbekanntes entdecken- verstanden. Vermuten wird als Synonym für ableiten verwendet.
		anig2508 Die Formel hat den Sinn, dass, wenn man sie ausrechnet, da auch über die Menge an Luft, ja, herausfindet.	Das Ausrechnen der Luftmenge in der Lunge ist eine Vorhersage über die Menge an Luft in der Lunge

- Der Proband hat die Aussage nicht auf dem intendierten Niveau verstanden.

Niveau	**Beispiel Antwort-alternative**	**Schüler*innenaussage**	**Begründung/Regel**
I	… die menschlichen Abstammungsverhältnisse vereinfacht abzubilden.	doon1101: man kann an den Schädeln ziemlich gut nachweisen, zum Beispiel, also anhand der Schädelähnlichkeit, dass der Mensch vom Affen abstammt.	„abbilden“ ≠ „nachweisen“
II	… die Funktionen der verschiedenen Gehirnareale begreiflich zu machen.	bela1402: das ist für mich wieder ein bisschen dasselbe wie die Funktionen vorauszusagen.	„begreiflich machen“ wird nicht als „erklären“, sondern auf Niveau III verstanden als Synonym zu „voraussagen“
	… die Abstammungs-verhältnisse menschlicher Vorfahren zu erklären.	inim2712: hier stehen jeweils die Namen der einzelnen Menschenarten und die werden ja dargestellt.	„erklären“ wird auf Niveau I verstanden als Synonym zu „darstellen“
III	… die Funktionen der einzelnen Gehirnareale vorauszusagen.	chea0611: Da habe ich mich für "gar nicht" entschieden, weil man an diesem Modell ja nicht erkennt, was die Funktion ist, sondern nur, wo es ist.	Das Modell wird nicht als Werkzeug für Voraussagen gesehen, sondern als Hilfsmittel, um etwas Bestehendes nachzuvollziehen.

Attachment 8. Instructions for the survey procedure for study III.

Ablauf Befragung

Phase	Ablauf		Material
Einführung	Sarah erklärt Ablauf der Befragung		Einführungsworte
Offene Aufgaben	SuS bearbeiten offene Aufgaben		Fragebogen 1 - Offen
Interview	Erste SuS, die fertig sind mit dem Bogen, werden zu einer kurzen mündlichen Befragung hinaus gebeten.	• Bedanken für Bereitschaft • Wir werden kurzes Gespräch führen • Darf Gespräch aufgezeichnet werden? • Passwort auf Band sprechen • Einstiegsimpuls, Fragen Wenn wieder zurück im Raum, dem Probanden den FC Fragebogen (2) geben.	Fragebogen mitnehmen Aufnahmegerät Laminierte Fragen
FC Aufgaben	SuS bearbeiten FC Aufgaben		Fragebogen 2 - FC

Du hast grade zwei Fragen zu Modellen beantwortet. Ich werde dir gleich nochmal zwei Fragen stellen und ich würde dich bitten, mir mündlich nochmal ein bisschen genauer zu beschreiben, was du denkst.

Beschreibe, inwieweit ein von Biologen entwickeltes Modell seinem biologischen Original entspricht.

Beschreibe, inwieweit sich ein von Biologen entwickeltes Modell vom biologischen Original unterscheidet.

Wenn wenig gesagt wird oder auf die Beantwortung der ersten Frage verwiesen wird, bitte sagen:

- Sag es ruhig noch einmal.
- Versuche das nochmal an einem Beispiel zu erklären.

Attachment 9. Categories (left) and examples (right) from the coding scheme for the open-ended justification tasks for the aspects ‘nature of models’ (T. rex – level III) and ‘purpose of models’ (T. rex – Level III).

Nature of models	
T. rex – level III: „The model of the T. rex shows the T. rex as it may have looked like but the scientists only presume how it really looked like.”	
Understood	*The bones have been put together like a puzzle. Not randomly but that doesn’t mean T. rex looked like that. There have been false assumptions in biology before, thus I adopt a distance to the correctness of the model.*
Not understood	*Based on the bones, the scientists can properly depict the physical build of the T. rex but they don’t know about the color.*
Strategy	*Randomly selected!*
Not ratable	*In the past, T. rex was presented differently by scientists.*
Purpose of models	
T. rex – level III: “The model of the T. rex permits to derive assumptions about the way that T. rex moved.”	
Understood	*I saw it on TV once where they made models of the T. rex and then they thought about how the model could move.*
Not understood	*Because the T. rex is shown in a pose that permits me to deduce how exactly it moved.*
Strategy	*I have decided to pick this answer because it is the only one that seems serious to me.*
Not ratable	*I would have made the model for that reason.*

Attachment 10. Coding scheme for the thematic qualitative content analysis (study VII).

(1) Die Schüleraussagen in den Interviews werden den vorgegebenen Kategorien/Codes zugeordnet.
(2) Bei der Codierung werden pro Schüleraussage minimal ganze Sätze oder – falls dies für die Nachvollziehbarkeit der Zuordnung notwendig ist – maximal ganze Kontexteinheiten den Kategorien zugeordnet.
(3) Einzelne Sätze oder ganze Schüleraussagen können mehreren Kategorien zugeordnet werden.
(4) Bei Unsicherheiten bezüglich der Zuordnung oder diskussionswürdigen Auffälligkeiten werden die entsprechenden Textstellen im Text markiert und mit einem erklärenden Memo versehen.
(5) Sonderregelungen für die Teilkompetenzen siehe Tabellen. [next page]

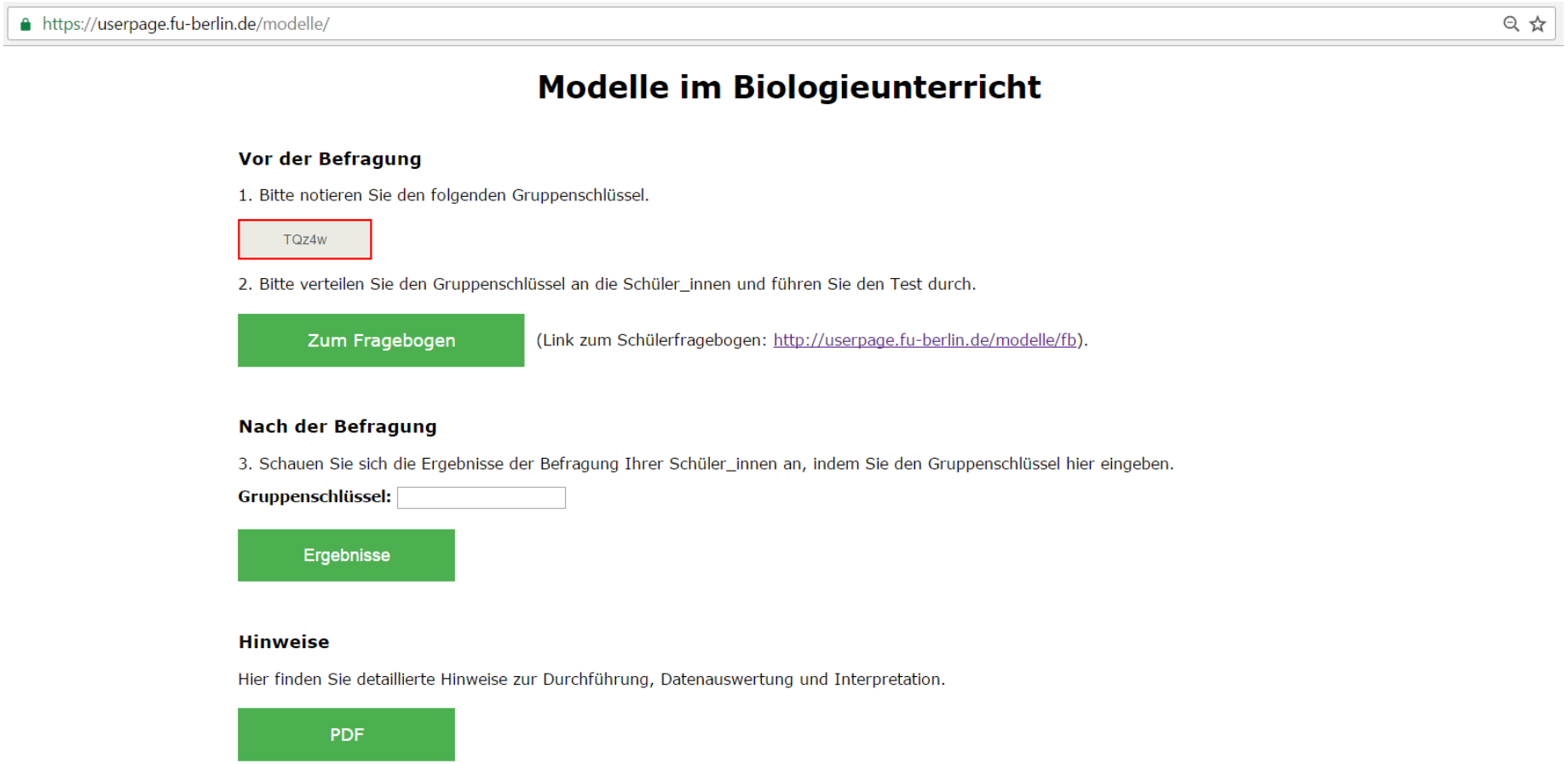
Sonderregelungen Teilkompetenz Eigenschaften von Modellen

Niveau I	
• **Model entspricht dem Fachwissen** Das Modell entspricht dem, was der Schüler über das Original weiß oder was er meint, das Konsenswissen über das Original ist.	*Ich kenne dieses Modell schon so seit der Grundschule und erachte es deswegen als richtig.* [sane1108] - Aussagen enthalten einen Bezug zum Wissen über das Original mit der Konnotation, dass es ein festes Wissen gibt. - S hat ein Bild vom Original im Kopf mit dem er das Modell vergleicht - (negativ) Aussagen enthalten ein Unwohlsein bzgl. der Richtigkeit des dargestellten Modells, weil das Original eigentlich anders ist
• **Wissenschaftler haben das herausgefunden** Hier wird die Arbeit von Forschern als Ursache für die Richtigkeit des Modells gegeben.	*Wissenschaftler haben nach langen Forschungen herausgefunden, wie der Wasserkreislauf genau funktioniert.* [ulia0610] - Wörter wie Wissenschaft, erforscht, Wissenschaftler, Forscher, herausgefunden, belegt… sind ein Hinweis - Das im Modell dargestellte Wissen scheint ein Endprodukt einer Tätigkeit zu sein
• **Belege durch Technik oder Daten** Das Modell entstand auf der Basis von Funden oder Untersuchungen.	*Man kann aufgrund heutiger moderner Techniken jedes Detail der Biomembran sehen.* [suie1710] - S erwähnt Technik (Mikroskope, Messungen usw.) oder Daten (Aufzeichnungen, Funde, Beobachtungen usw.) als Grundlage für die Erstellung des Modells
Niveau II	
• **Modell muss nicht alles zeigen** Das Modell ist auf irgendeine Weise vereinfacht und das ist auch sinnvoll.	*…n gab es bestimmt davor noch mehr Vorfahren. Da haben die Biologen gesagt, das können wir rauslassen, weil vielleicht ist das nicht mehr menschenähnlich.* [mama0503] - Bestimmte Teile sind nicht so wichtig, einige Teile sind wichtiger als andere
• **Nicht möglich, Modell gleich zu machen** Die Modellbauer sind nicht in der Lage 100%ige Übereinstimmung zum Original zu erreichen	*…ischen den abgebildeten Schritten sind noch kleinere Stufen, die in einem Modell nicht gezeigt werden können.* [coja2101] - Es scheint nicht möglich zu sein, alle Details des Originals im Modell darzustellen. - Nicht unbedingt, weil sie noch nicht erforscht sind, sondern weil die Bauer des Modells das nicht können.
• **Originale können sich unterscheiden** Das Modell muss verallgemeinert sein, weil sich die Originale horizontal oder chronologisch voneinander unterscheiden.	*In der Natur kann es immer Abweichungen geben und das ist ein Modell. Es ist ein idealisierter Grippe Virus.* [anra2109] - Die Natur verändert sich, Mutationen, Umweltveränderungen usw. - Die Natur ist vielfältig und ein Individuum gleicht nicht dem nächsten.
• **Teile sind bekannt, andere fehlen** S macht allgeimeine Aussagen darüber, dass Informationen fehlen, warum genau wird nicht spezifiziert.	*Wissenschaftler sind in der Lage, einige Informationen zu ermitteln, andere Informationen wurden jedoch nach einer Theorie/Hypothese festgelegt.* [jaca1811] - Man hat nicht alle Informationen. - Wenn man Infos nicht kann, kann man sie auch nicht im Modell berücksichtigen.
• **Fehlende Informationen können nicht gefunden werden**	*Einige Körperteile, die nur aus Fleisch bestanden, können heute nicht mehr nachgewiesen werden.* [dual1405] - Wie oben fehlen Informationen. - Grund besteht darin, dass das Original für den Meschen nicht zugänglich ist (z. B. ausgestorben usw.).
• **Fehlende Informationen wurden NOCH nicht gefunden**	*…odelle sind unvollständig. Vielleicht gibt es Dinge an einer Biomembran, die noch nicht entdeckt wurden. Die Wissenschaftler haben sie nicht eingebaut, weil sie (noch) nicht wissen, dass es sie gibt.* [jrld1809] - Wie oben fehlen Informationen. - Grund besteht darin, dass noch nciht genug geforscht wurde, Technik noch nicht so weit ist usw. - Schlagwörter sind “noch nicht”, “schon” oder Aussagen über die Zukunft
Niveau III	
• **Unsicherheit des Wissens bleibt bestehen** S meint, dass man sich insgesamt immer unsicher ist/sein wird über das Original.	*…eil der T. rex vor so langer Zeit lebte. Natürlich kann er kurze Arme, ein großes Gebiss oder diese Körperformen aufgewiesen haben. Es sind aber alles nur Vermutungen. Die nicht mehr und niemals bewiesen werden können.* [kite1511] - Es geht hier nicht um einzelne fehlende Aspekte, sondern um generelle Aussagen über das ganze Original. - “Wir haben zu der Zeit nicht gelebt”, “Es sind nur Theorien”, “Es wird niemals bewiesen werden” usw.
• **Unsicherheit des Modells bleibt bestehen**	*…odelle sind Erzeugnisse aus Forschungen und diese kann man aufzeigen, wie man behauptet, dass es ist bis einer das widerlegt.* [anma2012] - Wie bei Oberkategorie - Klarer Bezug zum Modell als Wissensform
• **Wissenschaftler können sich irren** Wissenschaftler können nur Informationen, die es gibt, deuten.	*…f dem Mikroskop-Bild sind Formen zu sehen, die im Modell von Wissenschaftlern gedeutet wurden, welche aber auch anders deutbar sind.* [vein1809] - Das Modell wurde vielleicht auch schon einmal anders dargestellt - Vermutung der Wissenschaftler steht im Vordergrund

Sonderregelungen Teilkompetenz Zweck von Modellen

Niveau I	
• **Das Original ist im Modell sichtbar** Das Modell zeigt Teile oder das ganze Original, so wie es in der Antwortalternative beschrieben ist.	*Ich habe mich für diese Antwort entschieden, weil man den Aufbau erkennen kann.* [anie0910] - S beschreibt Eigenschaften des Originals, die im Modell gut zu erkennen/wahrzunehmen/sehen/sichtbar sind - Antworten beziehen sich deskriptiv auf das im Modell dargestellte Original.
• **Sich selbst ein Bild machen** Das Modell hilft einem dabei, sich selbst ein Bild vom Original zu machen.	*Da ich mir direkt ein Bild von der Fortbewegung des T. rex machen konnte.* [page2210] - Das eigene Lernen steht im Vordergrund; man macht sich das Wissen zu eigen. - Passiv-Aussagen (man usw.), persönliche Aussagen (ich, für mich, usw.)
• **Anderen ein Bild machen** Das Modell hilft dabei, anderen etwas über das Original zu kommunizieren.	*Ich wählte diese Antwort, weil ich glaube, dass man den Leuten zeigen will, wie ein T. rex so aussah.* [reko2711] - Kommunikation steht im Mittelpunkt; das Wissen soll zugänglich gemacht werden.
Niveau II	
• **Modell wurde im Unterricht genutzt** Das Modell ist didaktisch wertvoll.	*Das hatten wir letztes Jahr. Damit hat unsere Lehrerin versucht, uns den Aufbau der Biomembran zu erklären.* [elsa2897] - S stellt bezug zum Unterricht als Quelle seines Wissens dar - Didaktische Eignung in der Vergangenheit oder Zukunft wird erwähnt.
• **Modell ist vereinfacht** Das Modell ist auf irgendeine Weise vereinfacht und das ist auch sinnvoll.	*n man anhand solcher Darstellungen vieles deutlicher machen kann und somit auch besser erklären kann.* [stna1904] - Modell eignet sich für Erklärungen, weil es einfach, übersichtlich, klar, deutlich usw. Ist.
• **Variablen lassen sich sinnvoll in Beziehung setzen**	*an sieht deutlich, dass der T. rex hohe Geschwindigkeiten erreichen konnte, da die zwei langen Beine problemlos weitere Schritte hinter sich bringen konnten.* [mosa2304] - Es wird über Zusammenhang/Einfluss/Abhängigkeit, über eine Kausalität oder eine Korrelation gesprochen.
~~**Niveau II**~~	
• **Erklärungen bedürfen Beschriftungen** Das Modell eigenet sich nicht für Erklärungen, weil ihm Text fehlt.	*Dieses Modell "erklärt" mir nichts. Erklärungen hängen für mich immer an einem Text.* [kaie2207] - Beschriftungen/Markierungen/Begriffe/Texte/Erläuterungen/weitere Angaben fehlen. - Texte sind besser zur Erklärung geeignet.
Niveau III	
• **Modell ist Ausgangspunkt für Hypothesen und Untersuchungen** Mit dem Modell selbst kann weiter geforscht warden.	*eil man dieses Modell auf dem Computer laufen lassen kann und so die Bewegungsabläufe rekonstruieren kann.* [yvck0311] - Der Blick richtet sich in die Zukunft. - Wörter wie: vermuten, spekulieren, weiter forschen, entdecken, weiterführen, unerforscht usw.
~~**Niveau III**~~	
• **Modell ist Endprodukt der Forschung** Modell stellt Wissen dar.	*dem Modell wird nicht weiter geforscht, da ein Modell nur das zeigen kann, was schon erforscht ist.* [yvle0406] - Modell kann nur ziegen, was schon erforscht ist – retrospektiver Blick. - S versuchen vergeblich, im Modell die Vorhersage abzulesen. - Verneinung von Verben: vorhersagen, vermuten, weiter forschen usw.
• **Modell ist zu einfach für Forschung** Forschung ist complex.	*n weiteres Erforschen ist nicht möglich, da nur ein Ausschnitt zu sehen ist und nicht alle Details dargestellt sind.* [keel1102] - Modell enthält nicht genug Details/Informationen, ist zu simple. - Es gibt bessere Modelle zu Forschung. - Forschungs ist komplex, kompliziert.
• **Modell ist Momentaufnahme** Modell ist für Objekte passend nicht für Prozesse.	*twicklung und Wachstum müssen über einen Zeitraum beobachtet werden und für mich sind Modelle nur eine Momentbetrachtung, an der man den Aufbau vieler Dinge zeigt.* [chne1506] - Bei Entwicklungen braucht man etwas, dass Entwicklungen darstellen kann wie einen Film oder eine Animation. - Schülern fehlt die Vorstellungskraft, dass es sich nicht um ein Bild handelt, sondern das Bild das Modell nur repräsentiert.
• **Forschung lieber am Original** Modell eignet sich nicht zum Forschen.	*n Modell dient zur Veranschaulichung etc. und nicht zum Forschen, dafür nimmt man das Original.* [bren2105] - Modell als Hypothesenbasis nicht erkannt. - Abstraktion nicht bewusst, Forschung muss unter genau den gleichen Bedingungen ablaufen wie beim Original.

Attachment 11. Screenshot of the online platform.



Attachment 12. Php script for the regression analysis. Script by Christoph van Heteren-Frese. Explaining remarcs in italics added by Sarah Gogolin.

```
++++++++++++++++++++++++++++++++++++++++++++++++++++++++++++++++++++++++++++++
        Datei: danke.php
        Inhalt: Berechnung des erreichten Niveaus
++++++++++++++++++++++++++++++++++++++++++++++++++++++++++++++++++++++++++++++

-->
<!-- PHP-Funktionen einbinden:-->
<?php
        include("./function.inc.php");
        include("./berechne_regression.php");

// Get data from databas
        $daten_Eigenschaften = hole_daten(["tre","vse","eve","nte","bme","wke"]);
        $daten_Zweck = hole_daten(["jwz","evz","ufz","bmz","ghz","trz"]);

// Start new regression for ‚Nature‘
// Coefficient matrix expressed as: b = (X'X)-1X'Y
        $reg_Eigenschaften = new Regression();[50]
        $reg_Eigenschaften->setX($daten_Eigenschaften[1]);
        $reg_Eigenschaften->setY($daten_Eigenschaften[0]);
        $reg_Eigenschaften->Compute();

// Start new regression for ‚Purpose‘
// Coefficient matrix expressed as: b = (X'X)-1X'Y
        $reg_Zweck = new Regression();
        $reg_Zweck->setX($daten_Zweck[1]);
        $reg_Zweck->setY($daten_Zweck[0]);
        $reg_Zweck->Compute();

// Results of the estimation
// Multiple regression expressed as: y = a+bx1+cx2+...
        $coefficients = $reg_Eigenschaften->getCoefficients();
        $coefficients_Zweck = $reg_Zweck->getCoefficients();
        $ergE=$coefficients[6]*$_POST["wke"]+$coefficients[1]*$_POST["tre"]+
                $coefficients[2]*$_POST["vse"]+$coefficients[3]*$_POST["eve"]+
                $coefficients[4]*$_POST["nte"]+$coefficients[5]*$_POST["bme"]+$coefficients[0];
        $ergZ=$coefficients[6]*$_POST["jwz"]+$coefficients[1]*$_POST["evz"]+
                $coefficients[2]*$_POST["ufz"]+$coefficients[3]*$_POST["bmz"]+
                $coefficients[4]*$_POST["ghz"]+$coefficients[5]*$_POST["trz"]+$coefficients[0];

// Format numbers, no decimal places
        $ergE=number_format($ergE,0);
        $ergZ=number_format($ergZ,0);

// Write results into _POST-Variable
        $_POST["nE"]=$ergE;
        $_POST["nZ"]=$ergZ;
?>
<!-- Übernommene Angaben des Probanden an DB senden -->
<?php sende_daten() ?>
```

[50] include("./Matrix.php"); namespace mnshankar\LinearRegression; Copyright (c) 2011 Shankar Manamalkav <nshankar@ufl.edu>; https://mnshankar.wordpress.com/2011/05/01/regression-analysis-using-php/

Attachment 13. ‘Information for teachers’ pdf on the userpage.

Hinweise zur Durchführung, Datenauswertung und Interpretation

Freie Universität Berlin

Mithilfe des Online-Fragebogens diagnostizieren Sie schnell und zuverlässig das Modellverstehen Ihrer Schüler_innen. Dank automatischer Datenauswertung und individualisierter tabellarischer Darstellung können Sie direkt nach der Diagnose auf die Ergebnisse Ihrer Klasse zugreifen. Durch Wiederholen des Fragebogens in einem Nachtest, können Sie den Fördererfolg evaluieren.

Befragen Sie Ihre Schüler_innen schnell und zuverlässig

- ✓ Die Schüler_innen beantworten an einem Computer mit Internet je sechs Aufgaben zu den Eigenschaften und dem Zweck von Modellen (ca. 10-15 min).
- ✓ In den Aufgaben zu verschiedenen biologischen Modellen (z. B. Wasserkreislauf oder Biomembran) wählen die Schüler_innen diejenige Perspektive aus, die ihrer eigenen Meinung am ehesten entspricht.
- ✓ Die angebotenen Perspektiven in den Aufgaben entsprechen den Niveaus des Kompetenzmodells der Modellkompetenz.

Die Daten werden automatisch für Sie ausgewertet

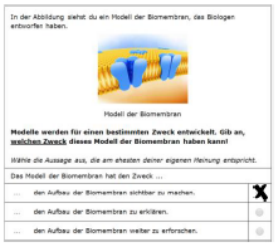
- ✓ Aus den insgesamt zwölf Antworten werden pro Schüler_in zwei Globalniveaus (1x für Eigenschaften und 1x für Zweck von Modellen) berechnet.
- ✓ Die Berechnung des Globalniveaus erfolgt durch Regressionsanalysen, die unterschiedliche Schwierigkeiten der Aufgaben berücksichtigen.
- ✓ Die Beschreibungen der Globalniveaus entsprechen den Niveaus des Kompetenzmodells der Modellkompetenz.

Fördern Sie Ihre Schüler_innen individualisert durch eine einfache Ergebnisinterpretation

- ✓ Nach der Bearbeitung erhält jede Schüler_in eine direkte Rückmeldung in Form eines zweigeteilten Kreises (Farblegende siehe Tabellen).
- ✓ Die linke Hälfte des Kreises steht für das Globalniveau in der Teilkompetenz Eigenschaften und die rechte Hälfte für Zweck von Modellen.
- ✓ Über Eingabe des Gruppenschlüssels auf der Homepage können die Ergebnisse der ganzen Klasse auf Aufgabenebene eingesehen werden.

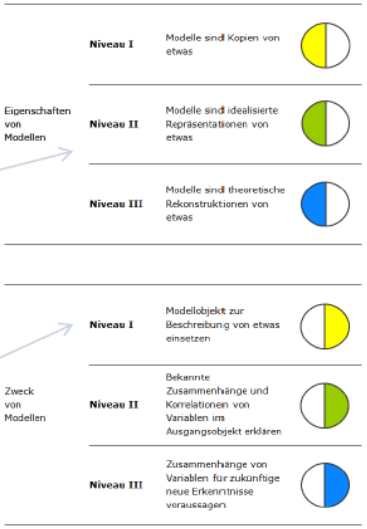
Aufgabenbeispiel: Eine von sechs Aufgaben zur Teilkompetenz Eigenschaften von Modellen

Bei dieser Aufgabe wurde eine Antwort auf Niveau II gewählt.



Aufgabenbeispiel: Eine von sechs Aufgaben zur Teilkompetenz Zweck von Modellen

Bei dieser Aufgabe wurde eine Antwort auf Niveau I gewählt.



Ergebnisbeispiel: Diese Schüler_in antwortet auf Niveau II in der Teilkompetenz Eigenschaften von Modellen und auf Niveau I in der Teilkompetenz Zweck von Modellen.

Nach Eingabe des Gruppenschlüssels erhalten Sie folgende Tabelle:

Ergebnisse der Gruppe *CFvOA*

Datum/Uhrzeit	Code	tre	vse	eve	nte	bme	wke	jwz	evz	ufz	bmz	ghz	trz	
2016-11-14 09:43:24	saah1905	1	2	2	2	2	2	1	1	1	1	1	1	
2016-11-14 09:45:20	heka0901	3	3	3	3	3	3	2	2	2	2	2	2	
2016-11-14 09:46:46	elne1205	1	1	1	1	1	1	1	1	1	1	1	1	

Literatur

Fleige, J., Seegers, A., Upmeier zu Belzen, A., & Krüger, D. (Hrsg.) (2012a). *Modellkompetenz im Biologieunterricht 7-10*. Donauwörth: Auer.
Gogolin, S. & Krüger, D. (in Druck). Modellverstehen im Biologieunterricht diagnostizieren und fördern. *Der mathematische und naturwissenschaftliche Unterricht*.
Grünkorn, J., Lotz, A., & Terzer, E. (2014). Erfassung von Modellkompetenz im Biologieunterricht. *Der mathematische und naturwissenschaftliche Unterricht, 67*, 132-138.

Sie haben noch Fragen? Wir beraten Sie gern! Email: sarah.gogolin@fu-berlin.de

10. Articles

Article 1 — Gogolin, S., & Krüger, D. (2015). Nature of models - Entwicklung von Diagnoseaufgaben. In M. Hammann, J. Mayer, & N. Wellnitz (Eds.), *Lehr- und Lernforschung in der Biologiedidaktik* (6th ed., pp. 27–41). Innsbruck: Studienverlag.

Article 2 — Gogolin, S. & Krüger, D. (2016). Diagnosing Students' Understanding of the Nature of Models. *Research in Science Education, 47, 5,* 1127–1149, doi: 10.1007/s11165-016-9551-9.

Article 3 — Gogolin, S., & Krüger, D. (2016). Konstruktion von Diagnoseaufgaben zum Zweck von Modellen. *Biologie Lehren und Lernen – Zeitschrift für Didaktik der Biologie*, *1*(20), 44–62.

Article 4 — Gogolin, S. & Krüger, D. (in press). Students' understanding of the nature and purpose of models. *Journal of Research in Science Teaching*.

Article 5 — Mathesius, S. & Gogolin, S. (2017). Die Letzten werden die Ersten sein – Praktisches Modellieren von Planktonkörperformen. *Biologie 5-10*, *17*, 10-13.

Article 6 — Gogolin, S. & Krüger, D. (in press). Modellverstehen im Biologieunterricht diagnostizieren und fördern. *Der mathematische und naturwissenschaftliche Unterricht*.

Article 7 — Gogolin, S., Krell, M., Lange-Schubert, K., Hartinger, A., Upmeier zu Belzen, A., & Krüger, D. (2017). Erfassung von Modellkompetenz bei Grundschüler_innen. In H. Giest, A. Hartinger, & S. Tänzer (Eds.), *Vielperspektivität im Sachunterricht* (pp. 108–115). Bad Heilbrunn: Klinkhardt-Verlag.

For Articles 1, 2, 3, 4 and 6, all data collection and analysis was performed by the first author. The manuscripts for these articles were conceptually structured by the author and a complete draft was reviewed by the co-author before being submitted and revised according to the suggestions made by the journals reviewers.

Article 5 was conceptually structured by the first and second author. All ideas and concepts were developed and brought to paper collabaratively by the two authors in joint sessions of writing.

Article 7 is conceptually structured by the first and second author. The first draft of the introduction and the theory section were produced by the second author. The method, findings and discussion sections were initially drafted by the first author. Both authors iteratedly reviewed and revised the manuscript before sending it to the remaining co-authors who equally gave feedback which was then integrated by the first author. The data collection for the article was performed under the guidance of author 3 and her research team. Data analysis was performed by author one, two and six. The test instrument was developed jointly by all authors.

Article 1 Gogolin, S., & Krüger, D. (2015). Nature of models - Entwicklung von Diagnoseaufgaben. In M. Hammann, J. Mayer, & N. Wellnitz (Eds.), *Lehr- und Lernforschung in der Biologiedidaktik* (6th ed., pp. 27–41). Innsbruck: Studienverlag.

[Manuscript attached as archivable pre-print]

Sarah Gogolin / Dirk Krüger

Nature of models - Entwicklung von Diagnoseaufgaben

Zusammenfassung

Der Perspektivwechsel vom Modell als Medium hin zum Verständnis, dass Modelle auch als Mittel zur naturwissenschaftlichen Erkenntnisgewinnung eingesetzt werden können, ist Bestandteil einer elaborierten Modellkompetenz. Darauf basierend besteht das Ziel dieses Projektes darin, ein computerbasiertes Instrument zu entwickeln, das Modellverstehen individuell und effizient diagnostiziert, um darauf aufbauend Modellkompetenz differenziert fördern zu können. Dieser Beitrag beschreibt die Entwicklung von Diagnoseaufgaben am Beispiel der Teilkompetenz *Eigenschaften von Modellen*. Die Fragestellungen, welche auf der Grundlage der Ergebnisse einer Vorstudie ($N = 467$) untersucht werden, beziehen sich auf die Beschreibung des Modellverstehens von Schülern[1] und die Analyse der Diagnoseaufgaben im Hinblick auf die Konstruktion des Diagnoseinstruments. Die Ergebnisse zeigen, dass Schüler Modelle vorwiegend unter medialer Perspektive beurteilen. Die entwickelten Aufgaben können zur Diagnose von Modellverstehen genutzt werden.

Abstract

The change of perspective, from seeing a model as a means of representing an original to the awareness that models are tools for scientific inquiry, is an important part of model competence. This project's objective is to construct a computer-based instrument which diagnoses students' individual understanding of models efficiently, in order to generate suggestions to foster their competence. This article presents the design of diagnostic tasks as an example of the aspect *nature of models*. The research

[1] Aus Gründen der besseren Lesbarkeit wird das maskuline Genus generisch verwendet und bezeichnet ebenso das weibliche wie das männliche Geschlecht.

questions are discussed on the basis of the results of an empirical survey conducted with 467 students and are aimed at describing the students' understanding of models, as well as the analysis of the diagnostic tasks with regard to the construction of the final instrument. The findings indicate that the majority of students understand models as representations. The analysis reveals, furthermore, that the developed tasks can be used to diagnose students' understanding of the nature of models.

Einleitung

Die Messung und Förderung von Kompetenzen auf der Basis von Kompetenzmodellen gehören zu den Herausforderungen der modernen Bildungsforschung (Fleischer, Koeppen, Kenk, Klieme & Leutner, 2013). Die Ergebnisse internationaler Schulleistungsstudien zeigen allerdings, dass die gesetzten Ziele im Bereich der naturwissenschaftlichen Bildung nicht vollständig zufriedenstellend erfüllt werden (u. a. Prenzel et al., 2007; Pant et al., 2013).
Im Zuge der Kompetenzorientierung tritt die Vermittlung naturwissenschaftlicher Denk- und Arbeitsweisen immer mehr in den Fokus des Biologieunterrichts (KMK, 2005). Hierbei spielt das Modellieren als Methode naturwissenschaftlicher Erkenntnisgewinnung eine bedeutende Rolle (u. a. Gilbert, Boulter & Elmer, 2000; KMK, 2005; Oh & Oh, 2011).
Upmeier zu Belzen und Krüger (2010) haben ein empirisch evaluiertes Kompetenzmodell entwickelt (vgl. Grünkorn, Upmeier zu Belzen & Krüger, 2014; Krell, 2013; Terzer, 2013), welches Fähigkeiten, die beim Denken über und im Umgang mit Modellen von Bedeutung sind, in fünf Teilkompetenzen strukturiert und eine Grundlage zur Diagnose von Modellkompetenz bietet. Dabei drückt sich eine elaborierte Perspektive in der wissenschaftlichen Nutzung von Modellen als methodische Werkzeuge (Niveau III) aus und lässt sich abgrenzen von Perspektiven auf Modelle als Medien (Niveau I und II; vgl. Upmeier zu Belzen & Krüger, 2010).
Ausgehend von der Forderung nach handlungsrelevanten Rückmeldungen in Bezug auf den Kompetenzstand und die -entwicklung von Schülern (Fleischer et al., 2013) wird im vorgestellten Projekt auf der Grundlage des Kompetenzmodells ein computerbasiertes Instrument entwickelt, das es erlaubt, im Biologieunterricht Modellverstehen individuell und effizient zu diagnostizieren, um darauf aufbauend Modellkompetenz differenziert fördern zu können. Dieser Artikel fokussiert auf die Entwicklung, Pilotierung und Optimierung von Diagnoseaufgaben. Ferner werden Implikationen vorgestellt, die sich aus der Untersuchung für das Diagnoseinstrument ergeben.

Theoretischer Hintergrund

Modellklassifizierung

Modelle stehen als Instrumente zur Erkenntnisgewinnung im Zentrum wissenschaftlicher Forschung (u. a. Bailer-Jones, 2002; Giere, 1999; Van der Valk, Van Driel & De Vos, 2007). Sie sind das Ergebnis eines hypothesengeleiteten Modellierungsprozesses und erlauben die Ableitung weiterer Fragestellungen und Hypothesen (Upmeier zu Belzen & Krüger, 2010). Modelle ermöglichen es Wissenschaftlern, die Gültigkeit von Theorien zu beurteilen und Erkenntnisse über Originale zu gewinnen, indem auf der Grundlage von Analogien die Passung zwischen dem Modell, der Theorie und der Empirie beurteilt wird (Mahr, 2009; Terzer, 2013). Oh und Oh (2011) stellen wie bereits Mahr (2009) und Stachowiak (1973) fest, dass es keine einheitliche Definition für den Modellbegriff gibt. In der Literatur findet sich zudem eine Vielzahl an Klassifizierungsversuchen von Modellen. Hervorzuheben ist für diese Arbeit die Klassifizierung von Suckling, Suckling und Suckling (1978), die im Bereich der Chemie zwischen gegenständlichen und konzeptuell-symbolischen Modellen unterscheiden. Boulter und Buckley (2000) beziehen sich auf die Repräsentationsformen von Modellen und unterscheiden fünf Perspektiven, darunter *concrete*, *visual* und *mathematical*. Diesen wissenschaftlichen Modellklassifizierungen steht die Sicht von Schülern gegenüber. In einer Studie von Krell, Upmeier zu Belzen und Krüger (2014) klassifizierten die befragten Schüler Modelle in die Gruppen *gegenständlich*, *abstrakt* und *besonders*.

Modellkompetenz

Die Bildungsstandards für den Mittleren Schulabschluss im Fach Biologie greifen Modelle und das Modellieren in fünf von 13 Standards im Bereich der Erkenntnisgewinnung auf (KMK, 2005, E9 – E13). Empirische Studien deuten darauf hin, dass Schüler die Bedeutung von Modellen im naturwissenschaftlichen Erkenntnisprozess nur wenig reflektieren. In Interviewstudien zeigt sich, dass Schüler Modelle meistens als Medien verstehen, um Bekanntes darzustellen und zu veranschaulichen. Sie sehen Modelle als im Maßstab veränderte bzw. idealisierte Kopien der Realität. Sehr wenige Schüler geben an, dass ein Modell eine Vermutung über das Original darstellt (u. a. Grosslight, Unger, Jay & Smith, 1991; Trier & Upmeier zu Belzen, 2009). Auch Lehrkräfte beschreiben Modelle bevorzugt als Medien zur Veranschaulichung, während sie deren Rolle als Werkzeuge im wissenschaftlichen Erkenntnisprozess kaum wahrnehmen

(u. a. Crawford & Cullin, 2005; Justi & Gilbert, 2003; Treagust, Chittleborough & Mamiala, 2002; Van Driel & Verloop, 2002).
Das Kompetenzmodell nach Upmeier zu Belzen und Krüger (2010) strukturiert Modellkompetenz in fünf Teilkompetenzen.
Die Dimensionalität von Modellkompetenz wurde im Rahmen der Evaluation des Kompetenzmodells durch Krell (2013) und Terzer (2013) überprüft. Terzer (2013) berichtet zunächst eine eindimensionale Struktur von Modellkompetenz, relativiert jedoch ihre Ergebnisse und betont, „dass eine eindimensionale Lösung nicht optimal ist und mehrere Dimensionen angenommen werden sollten." (Terzer, 2013, S. 161). Die Ergebnisse der Studie von Krell (2013) stützen die Annahme einer fünfdimensionalen Struktur von Modellkompetenz. Krell (2013) sieht den Mehrwert eines mehrdimensionalen und damit differenzierten Ansatzes jedoch vor allem in der didaktischen Anwendungssituation (vgl. z. B. Crawford & Cullin, 2005; Justi & Gilbert, 2003). Für die Diagnose von Modellverstehen wird entsprechend der Empfehlung von Fleige, Seegers, Upmeier zu Belzen & Krüger (2012), für die Förderung von Modellkompetenz im Biologieunterricht einzelne Teilkompetenzen getrennt zu fokussieren, das fünfdimensionale Modell zugrunde gelegt.
Jede der fünf Teilkompetenzen lässt sich in drei Niveaustufen unterteilen, die unterschiedlich elaborierte Perspektiven darstellen. Die Teilkompetenz *Eigenschaften von Modellen (nature of models)*, auf die in diesem Artikel der Blick gelenkt wird, deckt Perspektiven ab, die sich auf verschiedene Ähnlichkeits- bzw. Abstraktionsgrade zwischen Modell und Ausgangsobjekt beziehen (Upmeier zu Belzen & Krüger, 2010; Tab. 1).

Tab. 1: Kategoriensystem zur Teilkompetenz Eigenschaften von Modellen (nach Grünkorn et al., 2014)

Niveau I	Niveau II	Niveau III
Modelle sind Kopien von etwas	**Modelle sind idealisierte Repräsentationen von etwas**	**Modelle sind theoretische Rekonstruktionen von etwas**
Modell als Kopie	Modell ist in Teilen eine Kopie	Modell als hypothetische Darstellung
Modell mit großer Ähnlichkeit	Modell als eine mögliche Variante	
Modell entspricht (nicht) subjektiver Vorstellung vom Original	Modell als fokussierte Darstellung	

Schüler nehmen dabei ein Modell entweder als eine naturgetreue Replikation (Niveau I), eine idealisierte Repräsentation (Niveau II) oder eine theoretische Rekonstruktion (Niveau III) wahr (Tab. 1). Der Sprung vom Verständnis eines Modells als Abbild eines Originals hin zur Sicht auf ein Modell als Vermutung über ein Original erfolgt dabei von Niveau II zu Niveau III.

Vom Kompetenzmodell zum Diagnoseinstrument

Grünkorn et al. (2014; offenes Aufgabenformat) haben ausgehend von Schülerantworten in den Niveaus jeder Teilkompetenz Kategorien beschrieben und damit das Kompetenzmodell verfeinert (Tab. 1). Jede Kategorie in der Teilkompetenz *Eigenschaften von Modellen* repräsentiert dabei eine Perspektive, die von Schülern bezogen auf die Ähnlichkeit zwischen Modell und Ausgangsobjekt geäußert wurde.
Für die Evaluierung von Bildungsprozessen kommt der Diagnose von Kompetenzen eine Schlüsselfunktion zu (Hartig & Jude, 2007). Kompetenz wird von Klieme und Hartig (2007, S. 19) als die kontextspezifische „Verbindung von Wissen und Können in der Bewältigung von Handlungsanforderungen“ beschrieben. Die in diesem Projekt entwickelten Aufgaben erfassen die kognitiven Facetten der Modellkompetenz, während individuelle Bereitschaften und manuelle Fertigkeiten nicht erhoben werden. In Anlehnung an Krell (2013) erfassen die Aufgaben somit das Modellverstehen der Schüler, von dem auf Kompetenz geschlossen werden kann.
Grundsätzlich sollte hierbei eine mögliche schwierigkeitserzeugende Wirkung des Aufgabenkontexts berücksichtigt werden. Der Begriff Aufgabenkontext wird in diesem Zusammenhang im Sinne von Aufgabenstamm bzw. Aufgabenmerkmal verstanden. In Bezug auf das Modellverstehen von Schülern betonen z. B. Harrison und Treagust (2000) ebenso wie Krell et al. (2014), dass unterschiedliche Modelle jeweils spezifische kognitive Anforderungen transportieren. Dies sollte bei der Entwicklung eines Diagnoseinstruments überprüft und bei dessen Einsatz gegebenenfalls berücksichtigt werden (Nehm & Ha, 2011; Krell, Upmeier zu Belzen & Krüger, im Druck).
Mit dem offenen Aufgabenformat von Grünkorn et al. (2014) lassen sich Schülerperspektiven auf Modelle differenziert erheben. Gleichzeitig ist der damit verbundene hohe Auswertungsaufwand für effiziente Diagnosen unökonomisch und damit ungeeignet (Hartig & Jude, 2007).
Das Ziel dieses Projektes besteht darin, ein Instrument zur Individualdiagnose in den fünf Teilkompetenzen zu entwickeln. Ein solches Instrument soll in Schulen den Lehrkräften unmittelbar Hinweise über das Modellverstehen ihrer Schüler liefern und damit ermöglichen,

Modellkompetenz im Biologieunterricht individuell zu fördern. Ferner lässt sich mit einem solchen Instrument die Wirkung von Fördermaßnahmen ökonomisch evaluieren (Pant, 2013).

Fragestellungen und Hypothesen

Die für das Instrument entwickelten Diagnoseaufgaben erfassen beispielhaft in der Teilkompetenz *Eigenschaften von Modellen* das Modellverstehen von Schülern gemäß den im Kompetenzmodell (Upmeier zu Belzen & Krüger, 2010) dargestellten Ausprägungen. Hieraus ergibt sich die erste Forschungsfrage:

F1: Welches Niveau zeigen Schüler in der Teilkompetenz *Eigenschaften von Modellen*?

Grosslight et al. (1991), Grünkorn et al. (2014) sowie Trier und Upmeier zu Belzen (2009) beschreiben für qualitative Untersuchungen, dass Schüler Modelle vermehrt als Kopie eines Originals sehen (Niveau I). Krell (2013) und Treagust et al. (2002) zeigen in quantitativen Untersuchungen, dass die Mehrzahl der Schüler Modelle als idealisierte Repräsentationen eines Originals betrachtet (Niveau II).

Eine Analyse des Antwortverhaltens von Schülern bei unterschiedlichen Aufgabenkontexten bringt Aufschlüsse für die Konstruktion des Diagnoseinstruments:

F2: Inwieweit unterscheidet sich das Antwortverhalten der Schüler zwischen Aufgaben mit theoretisch als gleichartig einzustufenden Aufgabenkontexten?

Die in diesem Projekt theoretisch als gleichartig eingestuften Aufgabenkontexte unterscheiden sich sprachlich und inhaltlich nur geringfügig, weshalb bei solchen Aufgaben das Antwortverhalten der Schüler zu ähnlichen Ergebnissen führen sollte.

Methodisches Vorgehen

Entwicklung der Diagnoseaufgaben

Die Auswahl der Aufgabenkontexte basierte auf den Erfahrungen aus den bisherigen Forschungsprojekten mit offenen (Grünkorn et al., 2014), Multiple Choice- (Terzer, 2013) und Forced Choice-Aufgaben (Krell, 2013).

Für die Entwicklung von Aufgabenkontexten wurden Modelle ausgewählt, die sich in Anlehnung an Krell et al. (2014) zunächst in *gegenständlich* und *abstrakt* einteilen ließen (Tab. 2).

Tab. 2: Klasseneinteilung der Aufgabenkontexte
Die Aufgabenkontexte sind folgendermaßen abgekürzt: TR = *Tyrannosaurus rex*; NT = Neandertaler; BM = Biomembran; VS = Virus; EV = Evolution; WK = Wasserkreislauf; LS = Luftstrom; WM = Wassermelone.

Klassen	gegenständlich				abstrakt			
	Mesokosmos		Mikrokosmos		Schema		Formel	
Aufgaben-kontexte								
	TR	NT	BM	VS	EV	WK	LS	WM

Dabei beziehen sich zwei gegenständliche Aufgabenkontexte auf den Mesokosmos (TR, NT) und zwei auf den Mikrokosmos (BM, VS). Unter den abstrakten Aufgabenkontexten finden sich zwei, die Prozesse (EV, WK) schematisch darstellen, und zwei, die in Formeln spezifische Beziehungen (LS, WM) ausdrücken (Tab. 2). Allen Modellen ist gemein, dass sie sowohl aus medialer als auch aus methodischer Perspektive betrachtet werden können. In den jeweils drei Kategorien der Niveaus I und II der Teilkompetenz *Eigenschaften von Modellen* (vgl. Grünkorn et al., 2014; Tab. 1) wurden für die vier gegenständlichen Aufgabenkontexte (Tab. 2) je drei unterschiedliche Antwortalternativen entwickelt (d. h. eine Antwortalternative pro Kategorie). Drei inhaltlich identische Antwortalternativen wurden gemäß der einen Kategorie im Niveau III (Tab. 1) formuliert. Für die vier abstrakten Aufgabenkontexte (Tab. 2) wurden nur jeweils zwei Antwortalternativen pro Niveau entwickelt, da eine sinnvolle Formulierung von drei Alternativen nicht für alle Niveaus möglich war.
Die Antwortalternativen zu jeweils zwei theoretisch als gleichartig eingestuften Aufgabenkontexten wurden nahezu identisch formuliert (TR/NT, BM/VS, EV/WK und LS/WM; vgl. Tab. 2) und vor dem Einsatz durch Biologiedidaktik-Experten auf ihre Verständlichkeit und ihre Passung zum Niveau der Teilkompetenz überprüft.
Für die Pilotierung der Diagnoseaufgaben wurde jeweils eine Antwortalternative aus jedem Niveau in eine Rangordnungs-Aufgabe implementiert. Bei diesem Forced Choice-Aufgabentyp müssen die drei präsentierten Antwortalternativen in eine Rangfolge gebracht werden (Bock & Jones, 1968). Bei der Aufgabenentwicklung wurden alle Kombinationen der Antwortalternativen der drei Niveaus berücksichtigt.

Der Aufgabenkontext aller Aufgaben besteht aus grundlegenden Information zum biologischen Modell, Bildern zu Original und Modell und einer standardisierten Fragestellung (Abb. 1, vgl. Grünkorn et al., 2014; Krell & Krüger, 2010).

In der linken Abbildung siehst du den Knochenfund eines *T. rex* (*Tyrannosaurus rex*) und in der rechten Abbildung ein Modell eines *T. rex*, das Biologen entworfen haben.

Abbildung 1: Knochenfund eines *T. rex*.

Abbildung 2: Modell eines *T. rex*.

Gib an, inwieweit dieses Modell des *T. rex* so aussieht wie ein *T. rex*, der vor 68 Millionen Jahren lebte!

Schreibe die Buchstaben A, B oder C neben jede Aussage.

Das Modell des *T. rex* …	
… zeigt nur wesentliche Eigenschaften des Knochenfundes, z. B. die kurzen Arme, das große Gebiss und den langen Schwanz. (II)	B
… gleicht dem damals lebenden *T. rex* möglicherweise, sicher kann man aber nicht wissen, wie er ausgesehen hat. (III)	A
… stimmt mit dem damals lebenden *T. rex* überein, weil das Modell eine Nachbildung des Knochenfundes ist. (I)	C

Abb. 1: Forced Choice-Aufgabe zur Teilkompetenz Eigenschaften von Modellen

In der Instruktion wurden die Probanden gebeten, ihre Präferenz mit A (stärkste Zustimmung), B und C anzugeben. Die Angabe der Niveaus (I, II, III) dient hier der Illustration.

Pilotierung der Diagnoseaufgaben

Für die Befragung wurden die Aufgabenkontexte auf 24 verschiedene Testhefte verteilt. Zu den Kontexten TR, NT, BM und VS wurden je drei Forced Choice-Aufgaben, zu den Kontexten EV, WK, LS und WM je zwei Forced Choice-Aufgaben präsentiert. Dabei wurden in jeder Aufgabe drei Antwortalternativen (jeweils eine pro Niveaus) kombiniert. Die drei bzw. zwei Aufgaben pro Aufgabenkontext wurden den Schülern jeweils zusammenhängend präsentiert.
Jeder Schüler bearbeitete drei Aufgabenkontexte und fällte dabei pro Testheft (je nach Aufgabenkontexten) minimal sechs und maximal neun Entscheidungen. Die Befragung dauerte 20 Minuten und wurde mit 467 Schülern im Alter von 13 bis 18 Jahren durchgeführt.

In der Auswertung wurde jedem Schüler für jede Entscheidung das erstrangig gewählte Niveau (I – III) zugeordnet. Dies geschah in Anlehnung an Krell, Czeskleba und Krüger (2012), die zeigen konnten, dass bei entsprechenden Forced-Choice Aufgaben oft nur die erste Präferenz inhaltlich valide ausfällt. Zur Beantwortung der ersten Fragestellung wurden die Häufigkeitsverteilungen in den Niveaus analysiert.
Um Unterschiede im Antwortverhalten zwischen den einzelnen Aufgabenkontexten zu untersuchen (F2), wurden die Daten der theoretisch als gleichartig eingestuften Aufgabenkontexte verglichen (Mann-Whitney-U-Test). Für diesen Vergleich wurden nur die Daten derjenigen Schüler verwendet, die ein Testheft mit beiden Aufgabenkontexten bearbeitet hatten.

Ergebnisse

Modellverstehen der Schüler (Forschungsfrage 1)

Insgesamt präferieren 64,6 % der Schüler Aussagen auf Niveau I oder II und nehmen damit die präsentierten Modelle als Kopien oder idealisierte Repräsentationen eines Original wahr. Dagegen wählen 35,4 % der Schüler bevorzugt Aussagen auf Niveau III, die Modelle als theoretische Rekonstruktionen beschreiben (Tab. 3).

Unterschiede zwischen den Aufgabenkontexten (Forschungsfrage 2)

In den Aufgabenkontexten TR, NT und EV wird vermehrt das Niveau III präferiert, während in den anderen Aufgabenkontexten Niveau II bevorzugt gewählt wird (Tab. 3). Die Ergebnisse zeigen für die jeweils theoretisch gleichartig eingestuften Aufgabenkontexte mit Ausnahme der Paarung EV/WK ähnliche Häufigkeitsverteilungen (Tab. 3).

Tab. 3: Häufigkeiten der präferierten Niveaus in den einzelnen Aufgabenkontexten in Prozent.

	TR	NT	BM	VS	EV	WK	LS	WM	Gesamt
Niveau I	23,3	14,6	20,2	20,4	19,1	33,7	32,3	18,5	22,2
Niveau II	36,2	36,0	46,0	47,2	31,8	40,7	41,7	63,6	42,4
Niveau III	40,5	49,4	33,8	32,4	49,1	25,6	26,0	17,9	35,4

Die Ergebnisse des Mann-Whitney-U-Tests (Tab. 4) zeigen im Vergleich der jeweils theoretisch gleichartig eingestuften Aufgabenkontexte nur in

einem Fall einen signifikanten Unterschied. Im Aufgabenkontext Evolution (EV) präferieren die Schüler durchschnittlich ein signifikant höheres Niveau als im Aufgabenkontext Wasserkreislauf (WK) ($U = 5725$, $Z = -4{,}34$, $p < .001$). Die Effektstärke fällt klein aus ($r = .272$).

Tab. 4: Vergleich der theoretisch gleichartig eingestuften Aufgabenkontexte (Mann-Whitney-U-Test)

Aufgabenkontext		*N*		*Z*	*P* (2-seitig)
TR	*T. rex*	165	∑ 329	-1,819	n.s.
NT	Neandertaler	164			
BM	Biomembran	173	∑ 345	-0,958	n.s.
VS	Virus	172			
EV	Evolution	128	∑ 255	-4,340	$p < .001$
WK	Wasserkreislauf	127			
LS	Luftstrom	110	∑ 223	-0,680	n.s.
WM	Wassermelone	113			

Diskussion und Ausblick

Die Ergebnisse der Befragung deuten darauf hin, dass die entwickelten Aufgaben die Erfassung verschiedener Perspektiven auf Modelle zulassen und für eine Diagnose des Modellverstehens bei Schülern genutzt werden können. Die Beobachtung, dass die Mehrheit der Schüler Modelle unter medialer Perspektive beurteilt (Niveau I / II), stimmt mit den Befunden anderer Studien überein (Grosslight et al., 1991; Grünkorn et al., 2014; Krell, 2013; Trier & Upmeier zu Belzen, 2009). Ein möglicher Grund für die prominente Perspektive auf Modelle als Repräsentationen der Wirklichkeit ist nach Crawford und Cullin (2005) der Einfluss des Modellverstehens der Lehrkräfte auf die Vorstellungen der Schüler. Van Driel und Verloop (2002) gehen davon aus, dass Modelle im Unterricht primär als Mittel zur Veranschaulichung genutzt werden und ihr Potential zur Hypothesenbildung nur selten genutzt und zu wenig reflektiert wird. Die Ergebnisse der vorliegenden Studie unterstreichen die Notwendigkeit einer Förderung von Modellkompetenz im naturwissenschaftlichen Unterricht. Als Grundlage für eine differenzierte Förderung dieser Kompetenz wird im vorliegenden Projekt ein Diagnoseinstrument entwickelt.

Die Tatsache, dass beim Vergleich gleichartig eingestufter Aufgabenkontexte keine durchgängig und systematisch wiederkehrenden Unterschiede in der Verteilung festgestellt wurden, deutet auf die grundsätzliche Konsistenz der konstruierten Aufgaben hin. Die Verschiebung beim Aufgabenkontext Evolution (EV) gegenüber dem

Wasserkreislauf (WK) ins höhere Niveau ist möglicherweise auf den bekannten Status der Evolution als Theorie zurückzuführen. Dies würde erklären, weshalb die Mehrzahl der Schüler beim Aufgabenkontext Evolution (EV) Aussagen auf Niveau III präferieren.
Eine geeignete Strategie, die Gründe für Entscheidungen der Schüler empirisch zu überprüfen, stellt die Methode des „Lauten Denkens" (Sandmann, 2014) dar. Dabei lassen sich die Denk- und Entscheidungsprozesse der Schüler während der Aufgabenbearbeitung näher untersuchen und verstehen. Insgesamt bietet diese Methode auch die Möglichkeit zu prüfen, inwieweit der biologische Inhalt verschiedener Aufgabenkontexte als schwierigkeitserzeugendes Aufgabenmerkmal (Kauertz, 2008) wirkt. Da jedes Modell spezifische Anforderungen an einen Schüler stellt (Harrison & Treagust, 2000; Krell et al., 2014), sollte der Aufgabenkontext bei der Fällung eines Diagnoseurteils berücksichtigt werden.
Ferner ist eine Validierung der Aufgaben mittels halbstrukturierter Interviews (Niebert & Gropengießer, 2014) geplant, die im Anschluss an eine Aufgabenbearbeitung durchgeführt werden sollen. Hierbei kann außerdem festgestellt werden, wie sicher sich ein Schüler in seiner Perspektive auf ein bestimmtes Modell ist.
Die durch qualitative Untersuchungen gewonnenen Erkenntnisse über die Denkprozesse der Schüler und die spezifischen Anforderungen der einzelnen Aufgaben dienen zum einen dem Prozess der Validierung und geben zum anderen Hinweise darauf, wie die Beantwortung einer Aufgabe in einem computerbasierten Diagnoseinstrument gewertet und interpretiert werden soll. Schließlich soll ein adaptiver Computeralgorithmus entwickelt werden, der die aus der Vorstudie vorliegenden Wahrscheinlichkeitsverteilungen nutzt, um das Niveau eines neu befragten Schülers zu bestimmen. Das Verfahren soll zudem durch einen Algorithmus in der Lage sein, aus den Ergebnissen weiterer Befragungen zu lernen und somit die Diagnose stetig zu optimieren. Mit steigender Stichprobengröße können weitere diskriminierende Variablen wie z. B. der Aufgabenkontext oder die Klassenstufe des Schülers mit in die Diagnose einbezogen werden.
Unter einer Vermittlungsperspektive ist eine Reduktion der drei Niveaus des Kompetenzmodells (Upmeier zu Belzen & Krüger, 2010) auf zwei Perspektiven sinnvoll, wobei zwischen einer medialen (Niveau I / II) und einer methodischen Perspektive (Niveau III) auf Modelle unterschieden wird. Hierbei ergibt sich als Diagnoseurteil entweder die Empfehlung unterstützender Fördermaßnahmen von Schülern auf Niveau I / II oder einer expliziten und aktivierenden Herausforderung von Schülern auf Niveau III (vgl. Fleige et al., 2012; Upmeier zu Belzen & Krüger, 2013).

Literatur

Bailer-Jones, D. M. (2002). Naturwissenschaftliche Modelle: Von Epistemologie zu Ontologie. In A. Beckermann & C. Nimtz (Hrsg.), *Argument und Analyse – Sektionsvorträge* (S. 1–11). Paderborn: Mentis.

Bock, R., & Jones, L. (1968). *The measurement and prediction of judgment and choice*. San Diego, CA: Holden-Day.

Boulter, C. J., & Buckley, B. C. (2000). Constructing a typology of models for science education. In J. K. Gilbert & C. J. Boulter (Hrsg.), *Developing models in science education* (S. 41–57). Dordrecht: Kluwer Academic.

Crawford, B., & Cullin, M. (2005). Dynamic assessments of preservice teachers' knowledge of models and modelling. In K. Boersma, M. Goedhart, O. de Jong, & H. Eijkelhof (Hrsg.), *Research and the quality of science education* (S. 309–323). Dordrecht: Springer.

Fleige, J., Seegers, A., Upmeier zu Belzen, A., & Krüger, D. (2012). Förderung von Modellkompetenz im Biologieunterricht. *MNU*, *65*, 19–28.

Fleischer, J., Koeppen, K., Kenk, M., Klieme, E., & Leutner, D. (2013). Kompetenzmodellierung: Struktur, Konzepte und Forschungszugänge des DFG-Schwerpunktprogramms. *Zeitschrift für Erziehungswissenschaft, Sonderheft 18*, 5–22.

Giere, R. (1999). Modelle und Theorien. In V. Gadenne & A. Visintin (Hrsg.), *Wissenschaftsphilosophie* (S. 147–165). Freiburg: Karl Alber.

Gilbert, J. K., Boulter, C. J., & Elmer, R. (2000). Positioning models in science education and in design and technology education. In J. K. Gilbert & C. J. Boulter (Hrsg.), *Developing models in science education* (S. 3–17). Dordrecht: Kluwer Academic.

Grosslight, L., Unger, C., Jay, E., & Smith, C. (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.

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*. doi:10.1080/09500693.2013.873155

Harrison, A., & Treagust, D. (2000). A typology of school science models. *International Journal of Science Education*, *22*, 1011–1026.

Hartig, J., & Jude, N. (2007). Empirische Erfassung von Kompetenzen und psychometrische Kompetenzmodelle. In J. Hartig & E. Klieme (Hrsg.),

Möglichkeiten und Voraussetzungen technologiebasierter Kompetenzdiagnostik (S. 17–36). Bonn: BMBF.

Justi, R., & Gilbert, J. K. (2003). Teacher's views on the nature of models. *International Journal of Science Education*, *25*, 1369–1386.

Kauertz, A. (2008). *Schwierigkeitserzeugende Merkmale physikalischer Leistungstestaufgaben*. Berlin: Logos.

Klieme, E., & Hartig, J. (2007). Kompetenzkonzepte in den Sozialwissenschaften und im erziehungswissenschaftlichen Diskurs. *Zeitschrift für Erziehungswissenschaft*, *Sonderheft 8*, 11–29.

Klieme, E., Maag-Merki, K., & Hartig, J. (2007). Kompetenzbegriff und Bedeutung von Kompetenzen im Bildungswesen. In J. Hartig & E. Klieme (Hrsg.), *Möglichkeiten und Voraussetzungen technologiebasierter Kompetenzdiagnostik* (S. 5–15). Bonn: BMBF.

KMK (Sekretariat der Ständigen Konferenz der Kultusminister der Länder in der BRD, Hrsg.). (2005). *Bildungsstandards im Fach Biologie für den Mittleren Schulabschluss*. München: Wolters Kluwer.

Krell, M. (2013). *Wie Schülerinnen und Schüler biologische Modelle verstehen*. Berlin: Logos.

Krell, M., & Krüger, D. (2010). Diagnose von Modellkompetenz: Deduktive Konstruktion und Selektion von geschlossenen Items. *Erkenntnisweg Biologiedidaktik*, *9*, 23–37.

Krell, M., Czeskleba, A., & Krüger, D. (2012). Validierung von Forced Choice-Aufgaben durch Lautes Denken. *Erkenntnisweg Biologiedidaktik, 11*, 53–70.

Krell, M., Upmeier zu Belzen, A., & Krüger, D. (2014). How year 7 to year 10 students categorise models: Moving towards a student-based typology of biological models. In D. Krüger & M. Ekborg (Hrsg.), *Research in biological education* (S. 117–131). Verfügbar unter http://www.bcp.fu-berlin.de/biologie/arbeitsgruppen/didaktik/eridob_2012/eridob_proceeding/8-How-year.pdf?1389177404

Krell, M., Upmeier zu Belzen, A., & Krüger, D. (im Druck). Context-specificities in students' understanding of models and modelling in science: An issue of critical importance for both assessment and teaching. *Ebook proceedings of the ESERA 2013 conference*.

Mahr, B. (2009). Die Informatik und die Logik der Modelle. *Informatik Spektrum*, *32*, 228–249.

Nehm, R. H., & Ha, M. (2011). Item feature effects in evolution assessment. *Journal of Research in Science Teaching, 48*, 237–256.

Niebert, K., & Gropengießer, H. (2014). Leitfadengestützte Interviews. In D. Krüger, I. Parchmann & H. Schecker (Hrsg.), *Methoden in der*

naturwissenschaftsdidaktischen Forschung (S. 121–132) Berlin: Springer.

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

Pant, H. A. (2013). Wer hat einen Nutzen von Kompetenzmodellen? *Zeitschrift für Erziehungswissenschaft*, *Sonderheft 18*, 71–79.

Pant, H. A., Stanat, P., Schroeders, U., Roppelt, A., Siegle, T., & Pöhlmann, C. (Hrsg.). (2013). *IQB-Ländervergleich 2012: Mathematische und naturwissenschaftliche Kompetenzen am Ende der Sekundarstufe I*. Münster: Waxmann.

Prenzel, M., Schöps, K., Rönnebeck, S., Senkbeil, M., Walter, O., Carstensen, C. H., & Hammann, M. (2007). Naturwissenschaftliche Kompetenz im internationalen Vergleich. In M. Prenzel, C. Artelt, J. Baumert, W. Blum, M. Hammann, E. Klieme & R. Pekrun (Hrsg.), *PISA 2006. Die Ergebnisse der dritten internationalen Vergleichsstudie* (S. 63-105). Münster: Waxmann.

Sandmann, A. (2014). Lautes Denken – die Analyse von Denk-, Lern- und Problemlöseprozesse. In D. Krüger, I. Parchmann & H. Schecker (Hrsg.), *Methoden in der naturwissenschaftsdidaktischen Forschung* (S. 179–188). Berlin: Springer.

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

Suckling, C., Suckling, K., & Suckling, C. (1978). *Chemistry through models: Concepts and applications of modelling in chemical science and industry*. Cambridge, MA: Cambridge U.P.

Terzer, E. (2013). *Modellkompetenz im Kontext Biologieunterricht* (Dissertation). Humboldt Universität zu Berlin. Verfügbar unter http://edoc.hu-berlin.de/dissertationen/terzer-eva-2012-12-19/PDF/terzer.pdf

Treagust, D., Chittleborough, G., & Mamiala, T. (2002). Students' understanding of the role of scientific models in learning science. *International Journal of Science Education*, *24*, 357–368.

Trier, U., & Upmeier zu Belzen, A. (2009). „Wissenschaftler nutzen Modelle, um etwas Neues zu entdecken, und in der Schule lernt man einfach nur, dass es so ist.“ Schülervorstellungen zu Modellen. *Erkenntnisweg Biologiedidaktik*, *8*, 23–38.

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

Upmeier zu Belzen, A., & Krüger, D. (2013). Lernen mit Modellen. Modelle bauen und mit Modellen Neues erfahren. *Grundschule*, *6*, 6–9.

Van der Valk, T., Van Driel, J., & De Vos, W. (2007). Common characteristics of models in present-day scientific practice. *Research in Science Education*, *37*, 469–488.

Van Driel, J., & Verloop, N. (2002). Experienced teacher's knowledge of teaching and learning of models and modeling in science education. *International Journal of Science Education*, *24*, 1255–1277.

Article 2 Gogolin, S. & Krüger, D. (2016). Diagnosing Students' Understanding of the Nature of Models. *Research in Science Education, 1-23,* doi: 10.1007/s11165-016-9551-9.

[Article can be accessed via http://dx.doi.org/10.1007/s11165-016-9551-9]

Article 3

Gogolin, S., & Krüger, D. (2016). Konstruktion von Diagnoseaufgaben zum Zweck von Modellen. *Biologie Lehren und Lernen – Zeitschrift für Didaktik der Biologie*, *1*(20), 44–62.

[Article can be accessed via http://zdb.uni-bielefeld.de/index.php/zdb/article/view/328/pdf]

Sarah Gogolin & Dirk Krüger
Freie Universität Berlin

Konstruktion von Diagnoseaufgaben zum Zweck von Modellen

Development of diagnostic tasks for the purpose of models

Für die Entwicklung von Diagnoseinstrumenten wird in aktuellen Standards der Testentwicklung gefordert, die Schüler_innen stärker in den Prozess der Aufgabenentwicklung einzubeziehen. Dieser Beitrag schlägt eine auf Ratingaufgaben und Schülerinterviews basierende Prozedur zur Konstruktion von Forced Choice Aufgaben zum Zweck von Modellen vor. Die Konstruktionsschritte werden theoretisch begründet und an einem Beispiel illustriert. Die Prozedur ermöglichte es, sechs Diagnoseaufgaben zusammenzustellen, die eine theoriekonforme Interpretation des Modellverstehens von Schüler_innen ermöglichen.

Schlüsselwörter: Modelle, Biologie, Schülervorstellungen, Diagnose, Forced Choice Aufgaben, Validität

The latest standards for testing call for the integration of students into the process of test development. This article proposes a procedure for developing diagnostic tasks for the purpose of models which is both based on students' decisions on rating scales and on student interviews. The construction steps are being theoretically explained and justified as well as illustrated with an example. Using the procedure, six diagnostic tasks were developed which allow for an appropriate interpretation of students' understanding of the purpose of models.

Keywords: Models, Biology education, Students' understanding, Diagnosis, Forced Choice tasks, Validity

doi: 10.4119/UNIBI/zdb-v1-i20-328

1 Einleitung

Die *Standards for Educational and Psychological Testing* (AERA, APA & NCME, 2014) nennen Diagnose als eine von vielen Einsatzmöglichkeiten von Messinstrumenten in den Erziehungswissenschaften. Bei einer Diagnose wird die Erfassung und Interpretation von Merkmalsausprägungen verknüpft mit der Generierung handlungsrelevanter Rückmeldungen (Fleischer, Koeppen, Kenk, Klieme & Leutner, 2013). Der Einsatz von Messinstrumenten zur Diagnose ist traditionell jedoch mit wenigen Ausnahmen dem Bereich der Psychologie vorbehalten (Gorin, 2007). In den Erziehungswissenschaften dagegen werden Messinstrumente bislang häufig genutzt, um unter dem Paradigma der Outcomeorientierung Kompetenzen in large-scale Studien zu erfassen und zu beschreiben. Entsprechend wird in vielen Lehrbüchern der Erziehungs- und Sozialwissenschaften die Entwicklung von large-scale Messinstrumenten beschrieben (Gorin, 2007).

In den letzten Jahren stieg die Nachfrage nach einer Anwendung der vorhandenen, theoretisch beschriebenen Kompetenzmodelle für die schulische Praxis (Hartig, Klieme & Leutner, 2008). Entsprechend sollen Kompetenzausprägungen von Schüler_innen für die unterrichtliche Praxis nutzbar gemacht werden. Als Alternative zu large-scale Untersuchungen, die aufgrund der durch Multi-Matrix-Designs bedingten Messfehler nicht für Individualdiagnosen verwendet werden können (Kauertz, Neumann & Haertig, 2012), gilt es nun, effiziente Diagnoseinstrumente zu entwickeln, die individuelle Kompetenzausprägungen erfassen und differenzierte Ansatzpunkte zur Förderung liefern (Fleischer et al., 2013; Hartig et al., 2008). Unter methodischer Perspektive ergibt sich hieraus ein Bedarf an Prozeduren, die die Entwicklung aussagekräftiger Diagnoseinstrumente gewährleisten (Gorin, 2007; Leighton & Gierl, 2007). Dabei wird gefordert, die Schüler_innen stärker in den Prozess der Aufgabenentwicklung einzubeziehen (Adams & Wieman, 2011; Leighton & Gierl, 2007). Der National Research Council (NRC, 2001) macht das Übertragungspotential dieses Vorgehens auf Bildungskontexte in seinem Standardwerk zur Testentwicklung deutlich: „The methods used in cognitive science to design tasks […] are applicable to many of the challenges of designing effective educational assessments." (NRC, 2001, S. 5).

Dieser Beitrag stellt am Beispiel von Forced Choice Aufgaben zur Diagnose einer Teilkompetenz der Modellkompetenz, dem Zweck von Modellen, eine mehrschrittige Prozedur vor, die eine empirische Überprüfung der Aufgaben durch Schüler_innen mit einschließt. Diese Prozedur beruht auf den *Standards for Educational and Psychological Testing* (AERA et al., 2014) und erfüllt die vom NRC (2001) geforderten Voraussetzungen für die Entwicklung von Messinstrumenten (vgl. 2.2). In einem ersten Schritt werden ausgehend vom Kompetenzmodell der Modellkompetenz (Upmeier zu Belzen & Krüger, 2010) Antwortalternativen zum Zweck von verschiedenen biologischen Modellen konstruiert, die auf verschiedene Niveaus von Modellkompetenz und – innerhalb der Niveaus – auf verschiedene inhaltliche Aspekte des Modells fokussieren. Zur anschließenden Überprüfung der Antwortalternativen wird mittels kurzer Interviews untersucht, ob die Schüler_innen die Formulierungen wie intendiert verstehen. Auf der Basis von Ratingaufgaben wird überprüft, inwiefern die Antwortalternativen für Schüler_innen relevante inhaltliche Aspekte des jeweiligen Modells repräsentieren. Dieses Vorgehen dient der Beurteilung von Validität und

doi: 10.4119/UNIBI/zdb-v1-i20-328

wurde als Evidenzquelle genutzt, einerseits Antwortprozesse und andererseits den Testinhalt zu untersuchen (AERA et al. 2014). Abschließend werden geeignete Antwortalternativen in Forced Choice Aufgaben zusammengestellt.

2 Theoretischer Hintergrund

2.1 Zweck von Modellen

In der Wissenschaft werden Modelle als Hilfsmittel zur wissenschaftlichen Kommunikation und als Forschungsinstrumente genutzt (u. a. Harrison & Treagust, 2000; Passmore, Gouvea & Giere, 2014). Die Gründe, Modelle auch in der Schule zu nutzen, sind vielfältig. Modelle können zum einen bekannte Sachverhalte darstellen und sind damit Modelle *von* etwas (Mahr, 2008; Passmore, Gouvea & Giere, 2014). Damit eignen sie sich, um naturwissenschaftliches Wissen zu lernen (*learn science*; Hodson, 2014). Zum anderen können Modelle als Modelle *für* etwas gesehen werden (Mahr, 2008; Passmore, Gouvea & Giere, 2014), da aus Modellen Hypothesen abgeleitet werden, die mit Hilfe neuer Beobachtungen und Untersuchungen zur Konstruktion des erweiterten oder erneuerten Wissens führen (*learn to do science*; Hodson, 2014). Durch die Reflexion über die wissenschaftliche Arbeitsweise mit Modellen können Schüler_innen etwas über das Vorgehen in der Naturwissenschaft lernen (*learn about science*; Hodson, 2014).

Empirische Studien mit Schüler_innen zeigen, dass die Rolle von Modellen im wissenschaftlichen Erkenntnisprozess kaum wahrgenommen wird (u. a. Grosslight, Jay, Unger & Smith, 1991; Grünkorn, 2014; Treagust, Chittleborough & Mamiala, 2002; Trier & Upmeier zu Belzen, 2009). Grünkorn (2014) erfasste mittels offener Aufgaben Perspektiven, die Schüler_innen ($N = 706$) zum Zweck biologischer Modelle äußern. Hierbei gaben die befragten Schüler_innen an, der Zweck von Modellen sei entweder die Darstellung eines Sachverhalts oder das Erklären von Zusammenhängen. Weniger Schüler_innen äußerten, dass Modelle zum Überprüfen von Ideen genutzt werden könnten. Studien mit Lehrkräften naturwissenschaftlicher Fächer ergaben ebenfalls, dass Modelle im Unterricht vor allem als Medien zur Veranschaulichung eingesetzt werden (u. a. Crawford & Cullin, 2005; Justi & Gilbert, 2005; van Driel & Verloop, 2002). Trotzdem setzen Lehrkräfte ihrer eigenen Einschätzung nach Modelle im Unterricht auch wissenschaftlich für die Erkenntnisgewinnung ein (Krell & Krüger, 2013).

Nach dem Kompetenzmodell der Modellkompetenz (Upmeier zu Belzen & Krüger, 2010) werden fünf Teilkompetenzen (Eigenschaften von Modellen, Alternative Modelle, Zweck von Modellen, Testen von Modellen, Ändern von Modellen) unterschieden, die in drei Niveaus unterschiedliche Perspektiven auf Modelle beschreiben. In der Teilkompetenz „Zweck von Modellen“ (Tab. 1) bilden sich die zuvor genannten Schülerperspektiven ab. Niveaus I und II beschreiben einen vom Modellierer oder Modellnutzer definierten Anspruch an das Modell als Modell *von* etwas und in Niveau III den Anspruch an ein Modell als Modell *für* etwas (Upmeier zu Belzen & Krüger, 2010). Die Bildungsstandards (KMK, 2005) fordern explizit beide Perspektiven auf Modelle. Eine elaborierte Modellkompetenz zeigt sich dabei darin, dass Schüler_innen mit Modellen auch hypothesenbasiert im Sinne des Niveaus III argumentieren können.

doi: 10.4119/UNIBI/zdb-v1-i20-328

Tabelle 1: Niveaus der Teilkompetenz „Zweck von Modellen" (Upmeier zu Belzen & Krüger, 2010).

Niveau I	Niveau II	Niveau III
Modellobjekt zur Beschreibung von etwas einsetzen	Bekannte Zusammenhänge und Korrelationen von Variablen im Ausgangsobjekt erklären	Zusammenhänge von Variablen für zukünftige neue Erkenntnisse voraussagen

2.2 Entwicklung und Überprüfung von Diagnoseaufgaben

Eine individuelle Förderung von Schüler_innen bzgl. ihres Modellverstehens setzt dessen Diagnose voraus. Dabei sollen einerseits Vorstellungen zum Themenbereich ermittelt werden, die für Lehr- und Lernprozesse relevant sein können, und andererseits sollen Lernergebnisse festgestellt werden, die sich nach Vermittlungssituationen ergeben. Die Diagnose vor dem Lernprozess kann je nach Vorstellungs- oder Fähigkeitsprofil zu Zuweisungen der Schüler_innen zu Lerngruppen führen. In diesen Lerngruppen lässt sich individuelles Lernen mit angepassten Fördermaßnahmen optimieren (vgl. Ingenkamp & Lissmann, 2008).

Eine theoretische Basis für die Diagnose und Förderung der Fähigkeiten von Schüler_innen bilden Kompetenzmodelle. Diese beschreiben „Zusammenhänge zwischen individuellen Fähigkeiten und Fertigkeiten und erfolgreichem Handeln in spezifischen Kontexten" (Klieme & Hartig, 2007, S. 11). Explizite, handlungsrelevante Rückmeldungen an Lehrkräfte, wie beispielsweise die Zuweisung der Schüler_innen zu spezifischen Fördermaßnahmen, lassen sich aber nicht direkt aus den Modellen ableiten (vgl. Kauertz et al., 2012). Für diese Art der Rückmeldungen müssen individuelle Kompetenzausprägungen empirisch durch Diagnoseinstrumente bestimmt werden. Adams & Wieman (2011) betonen, dass Diagnoseinstrumente, die einen unterrichtlich nutzbaren Informationsgewinn schaffen, effizient ohne Training einsetzbar sowie valide interpretierbar sein müssen. Sie empfehlen hierfür den Einsatz von Aufgaben mit geschlossenem Antwortformat, da diese einfacher zu bearbeiten und auszuwerten sind. Für eine valide Interpretation der Diagnoseergebnisse ist bereits während der Aufgabenentwicklung sicherzustellen, dass (1) die Schüler_innen die Fähigkeit, die mit dem Verfahren gemessen werden soll, auch tatsächlich bei der Bearbeitung der Aufgaben einsetzen (Antwortprozesse; AERA et al., 2014) und (2) die Aufgaben das zu messende Konstrukt adäquat abbilden (Testinhalt; AERA et al., 2014).

Da Validität die Interpretation der Ergebnisse des Diagnoseinstruments betrifft, sollte das Verständnis der Diagnoseaufgaben empirisch mit der Stichprobe getestet werden, für die das Instrument konzipiert wurde (Antwortprozesse; AERA et al., 2014). Adams & Wieman (2011) schlagen vor, zum Zweck der Validierung während der Aufgabenentwicklung Schülerinterviews durchzuführen, in denen Schüler_innen ihre Antworten begründen oder erklären, wie sie bestimmte Begriffe verstehen. Sie verweisen in diesem Zusammenhang auch auf Ericsson und Simon (1998) und betonen: „Because these sorts of probing questions do alter student thinking and could likely help students think of connections they may not have in

doi: 10.4119/UNIBI/zdb-v1-i20-328

an actual testing situation, strict think-aloud interviews must be performed for validation once the test is constructed." (Adams & Wieman, 2011, S. 1297).

Um darüber hinaus Hinweise für Validität auf der Basis des Testinhalts zu erhalten, muss überprüft werden, ob die Aufgaben das zu messende Konstrukt adäquat abbilden (Testinhalt; AERA et al., 2014). Bei der Auswahl und Überprüfung von Aufgaben, die sich auf bestimmte Kontexte beziehen, sollte berücksichtigt werden, dass unterschiedliche Kontexte jeweils spezifische kognitive Anforderungen transportieren und dies einen Einfluss auf die Perspektiven von Schüler_innen haben kann (Krell, Upmeier zu Belzen & Krüger, 2014; Nehm & Ha, 2011). In der vorliegenden Studie wird die Kontextualisierung der Aufgaben durch die Verwendung unterschiedlicher biologischer Modelle in den Aufgabenstämmen realisiert. Der Begriff Kontext wird hier folglich im Sinne eines Item-Features genutzt (Krell, Upmeier zu Belzen & Krüger, 2012). Bei Aufgaben mit geschlossenem Antwortformat empfiehlt es sich, Antwortalternativen zu konstruieren, die die Perspektiven von Schüler_innen berücksichtigen (Haladyna, 2004). Hierbei können auf der einen Seite die Antwortalternativen auf der Grundlage zuvor erhobener Schüleraussagen konstruiert (Haladyna, 2004) und anschließend durch Experten in Bezug auf die Passung mit der zugrundeliegenden Theorie überprüft werden. Auf der anderen Seite kann nachträglich überprüft werden, ob die konstruierten und von Experten überprüften Antwortalternativen von den Schüler_innen tatsächlich als relevante Perspektiven in Bezug auf den Kontext der Aufgabe beurteilt werden (Leighton & Gierl, 2007).

3 Forschungsfragen und Hypothesen

Für eine effiziente und zugleich individuelle Diagnose von Modellverstehen sollen Aufgaben im Forced Choice Format eingesetzt werden, die je eine Antwortalternative pro Niveau zum Zweck von Modellen enthalten. Schüler_innen sollen aus den drei Antwortalternativen diejenige auswählen, die ihrer eigenen Meinung am ehesten entspricht (McCloy, Heggestad & Reeve, 2005). Für die Entwicklung solcher Diagnoseaufgaben wird zunächst ein Pool von Antwortalternativen für verschiedene biologische Kontexte konstruiert. Hierbei entstehen pro Niveau und Kontext mehrere Antwortalternativen, die auf verschiedene inhaltliche Aspekte des Kontexts fokussieren. Um die Antwortalternativen zu evaluieren, zu selektieren und eine Auswahl in Forced Choice Aufgaben zusammenzustellen, sind zwei Forschungsfragen empirisch zu untersuchen.

F1: Inwiefern verstehen die Schüler_innen die konstruierten Antwortalternativen auf dem jeweils theoretisch intendierten Niveau?

F2: Inwieweit unterscheiden sich die konstruierten Antwortalternativen in ihrer Relevanz für Schüler_innen in Bezug auf den jeweiligen biologischen Kontext?

doi: 10.4119/UNIBI/zdb-v1-i20-328

Es wird erwartet, dass die Schüler_innen die einzelnen Antwortalternativen entsprechend des theoretisch intendierten Niveaus verstehen, womit eine auf den Antwortprozessen basierende Evidenz für Validität erbracht wäre (AERA et al., 2014; Leighton & Gierl, 2007).
Um das Modellverstehen von Schüler_innen erfassen zu können, sollten – unabhängig vom Niveau – die in den Antwortalternativen fokussierten inhaltlichen Aspekte für Schüler_innen in Bezug auf den jeweiligen biologischen Kontext möglichst relevant sein. In Anlehnung an Studien, die inhaltliche Aspekte als schwierigkeitserzeugendes Aufgabenmerkmal betrachten, wird vermutet, dass der in den verschiedenen Antwortalternativen enthaltene Inhalt einen Einfluss darauf hat, wie relevant Schüler_innen die jeweilige Antwortalternative wahrnehmen (Cohors-Fresenborg, Sjuts & Sommer, 2004; Kauertz, 2008; Krell et al., 2014).

4 Methodisches Vorgehen

4.1 Entwicklung von Antwortalternativen

Als Ausgangspunkt für die Entwicklung der Forced Choice Aufgaben wurden für acht verschiedene biologische Kontexte je sechs Antwortalternativen konstruiert. Dabei wurden für jedes der drei Niveaus der Teilkompetenz „Zweck von Modellen“ (Tab. 1; Upmeier zu Belzen & Krüger, 2010) je zwei niveaugleiche Formulierungen entwickelt. Diese unterscheiden sich zum einen im niveaubestimmenden Verb, z. B. „sichtbar machen“ und „veranschaulichen“ als Indikatoren für Niveau I. Zum anderen beziehen sich die Antwortalternativen auf verschiedene Aspekte des in der Aufgabe genutzten Modells, z. B. auf den „Aufbau“ oder auf die „Bestandteile“. Nach ihrer empirischen Überprüfung sollte je eine Antwortalternative pro Niveau in eine Forced Choice Aufgabe implementiert werden (Tab. 2).
Es wurden Kontexte gewählt, die eine sinnvolle Formulierung von Antwortalternativen auf allen drei Niveaus der Teilkompetenz „Zweck von Modellen“ gemäß des Kompetenzmodells der Modellkompetenz erlauben.

Tabelle 2: Beispielhafte Darstellung der sechs entwickelten Antwortalternativen für den Kontext der Biomembran. Die niveaubestimmenden Verben sind fett gedruckt.

Das Modell der Biomembran hat den Zweck …	
Niveau I	… *den Aufbau der Biomembran* ***sichtbar zu machen.*** [Ia] … *die verschiedenen Bestandteile der Biomembran* ***zu veranschaulichen.*** [Ib]
Niveau II	… *den Aufbau der Biomembran* ***zu erklären.*** [IIa] … *das Zusammenwirken der Bestandteile* ***begreiflich zu machen.*** [IIb]
Niveau III	… *den Aufbau der Biomembran* ***weiter zu erforschen.*** [IIIa] … *weitere Bestandteile der Biomembran* ***vorauszusagen.*** [IIIb]

doi: 10.4119/UNIBI/zdb-v1-i20-328

4.2 Datenerhebung

Für die Untersuchung der Forschungsfragen und die empirische Überprüfung der Antwortalternativen unter Berücksichtigung der Einschätzungen von Schüler_innen wurden ein quantitativer und ein qualitativer Ansatz kombiniert (*convergent mixed method design*; Creswell & Plano Clark, 2011).
Für den quantitativen Ansatz wurden die jeweils sechs konstruierten Antwortalternativen pro Kontext in zufälliger Reihenfolge in Ratingaufgaben zusammengestellt (Abb. 1). Jede Aufgabe bestand aus einem standardisierten Aufgabenstamm, einer Abbildung des Modells sowie des Originals, einem standardisierten Impuls für die Aufgabenbearbeitung und den sechs Antwortalternativen.

In der linken Abbildung siehst du eine mikroskopische Aufnahme einer Biomembran und in der rechten Abbildung ein Modell der Biomembran, das Biologen entworfen haben.					
Abbildung 1: Mikroskopische Aufnahme einer Biomembran		Abbildung 2: Modell der Biomembran			
Modelle werden für einen bestimmten Zweck entwickelt. Gib an, <u>welchen Zweck</u> dieses Modell der Biomembran haben kann!					
Entscheide bei jeder Aussage, wie sehr sie deiner eigenen Meinung entspricht. Mache neben jeder Aussage nur ein Kreuz!					
Das Modell der Biomembran hat den Zweck ...	*gar nicht*	*wenig*	*teils-teils*	*annähernd*	*völlig*
... den Aufbau der Biomembran sichtbar zu machen. [Ia]	□	□	□	□	□
... die verschiedenen Bestandteile der Biomembran zu veranschaulichen. [Ib]	□	□	□	□	□
... den Aufbau der Biomembran zu erklären. [IIa]	□	□	□	□	□
... das Zusammenwirken der Bestandteile begreiflich zu machen. [IIb]	□	□	□	□	□
... den Aufbau der Biomembran weiter zu erforschen. [IIIa]	□	□	□	□	□
... weitere Bestandteile der Biomembran vorauszusagen. [IIIb]	□	□	□	□	□

Abbildung 1: Aufgabenbeispiel zum Kontext der Biomembran. Die Angabe der Antwortalternativen (Ia, Ib, IIa, ...) dient der Illustration. Die Reihenfolge der Antwortalternativen entspricht nicht der Reihenfolge im Fragebogen.

doi: 10.4119/UNIBI/zdb-v1-i20-328

Insgesamt bewerteten $N = 275$ Schüler_innen der Klassenstufen neun bis zwölf an Berliner Gymnasien auf einer fünfstufigen, verbaläquidistanten Likert-Skala (Abb. 1; vgl. Moosbrugger & Kelava, 2012), wie sehr die einzelnen Antwortalternativen ihrer eigenen Meinung entsprechen. Um die Arbeitsbelastung der Schüler_innen zu reduzieren, wurden die acht Aufgaben auf vier Testhefte aufgeteilt, die jeweils vier Aufgaben enthielten (*balanced incomplete block design*; Gonzalez & Rutkowski, 2010).
Für den qualitativen Ansatz wurden im Anschluss an die schriftliche Befragung mit je $n = 40$ Schüler_innen pro Klassenstufe vollstrukturierte Interviews geführt. Hierfür unterstützte der eigene, zuvor ausgefüllte Fragebogen als *Stimulated Recall* die Erinnerung an das Bearbeiten der Aufgaben (Sandmann, 2014). Im Interview lasen die Schüler_innen zunächst zur ersten Aufgabe die erste Antwortalternative und dann ihre Bewertung dieser auf der Likert-Skala vor. Anschließend wurden die Schüler_innen gebeten, ihre Bewertung zu begründen. Das gleiche Vorgehen wurde für alle sechs Antwortalternativen für zwei der vier im Fragebogen bearbeiteten Aufgaben durchgeführt. Insgesamt ergeben sich bei $n = 160$ interviewten Schüler_innen 1920 Schüleraussagen, folglich 40 Aussagen pro Antwortalternative.

4.3 Datenauswertung

Die Datenauswertung und die damit einhergehende sukzessive Selektion der Antwortalternativen verliefen zweigeteilt.
Um zu untersuchen, ob die Schüler_innen die einzelnen Antwortalternativen entsprechend des theoretisch intendierten Niveaus verstehen (Forschungsfrage 1), wurden die Audiodaten transkribiert und die Schüleraussagen mittels eines Leitfadens binär in „entsprechend der Theorie verstanden (1)“ oder „nicht entsprechend der Theorie verstanden (0)“ kodiert. Es war nicht von Interesse, ob das Niveau der Antwortalternative dem Niveau des Modellverstehens der Schüler_in entsprach. War aus der Schüleraussage nicht zu entnehmen, ob die Antwortalternative im Sinne des Niveaus verstanden wurde, wurde mit „8“ kodiert. 50 % der Aussagen wurden durch einen unabhängigen Rater zweitkodiert. Die Übereinstimmung der Rater wurde mittels Cohens Kappa überprüft (Wirtz & Caspar, 2002). Anschließend wurde ermittelt, wie viele Schüler_innen eine Antwortalternative entsprechend der Theorie verstanden haben. Es wurde zuvor in einer Expertenrunde (2 Wissenschaftler_innen aus dem Bereich der Biologiedidaktik, 2 Wissenschaftler_innen aus der Psychologie) festgelegt, diejenigen Antwortalternativen für eine weitere Verwendung in der Diagnose als geeignet einzustufen, die von mindestens 70 % der Schüler_innen entsprechend der Theorie verstanden wurden. Alle Antwortalternativen, die dieser Voraussetzung nicht genügten, wurden verworfen.
Verstanden die Schüler_innen beide Antwortalternativen eines Niveaus hinreichend gut, folgte ein weiterer Selektionsschritt, der die Relevanz der Antwortalternativen für die Schüler_innen (Forschungsfrage 2) nutzte. Hierfür wurden die Zustimmungen der Schüler_innen zu den zwei niveaugleichen Antwortalternativen auf der Likert-Skala (1-5; ca. 140 Bewertungen pro Antwortalternative) verglichen und damit auf die Relevanz der Antwortalternativen in Bezug auf den Kontext geschlossen. Für die Zusammenstellung der Forced Choice Aufgaben wurden diejenigen Antwortalternativen ausgewählt, die im

doi: 10.4119/UNIBI/zdb-v1-i20-328

Vergleich als relevanter in Bezug auf den angebotenen Kontext eingestuft wurden und somit am ehesten (vgl. McCloy et al., 2005) den Perspektiven von Schüler_innen beim jeweiligen Kontext entsprachen.

5 Ergebnisse

5.1 Passung zwischen intendiertem und interpretiertem Niveau

Die qualitative Bewertung der Schüleraussagen durch zwei unabhängige Rater wurde mit einer sehr guten Interrater-Reliabilität (Cohens-Kappa $0{,}86 < \kappa < 0{,}93$; Wirtz & Caspar, 2002) durchgeführt. Um die genutzten Kategorien vorzustellen, werden aus den Schüleraussagen ($N = 1920$) beispielhaft einige für die Kodierungen zu einer der beiden Antwortalternativen auf Niveau III für den Kontext der Biomembran vorgestellt (Tab. 3).

Tabelle 3: Beispielhafte Kodierung von Schüleraussagen zu einer Antwortalternative auf Niveau III beim Kontext der Biomembran.

Das Modell der Biomembran hat den Zweck, weitere Bestandteile der Biomembran vorauszusagen. [Antwortalternative IIIb]	
Kodierung	Ankerbeispiel
1 - Antwortalternative wurde entsprechend der theoretischen Intention verstanden	**Maon:** Weil ich dachte, da geht es darum, dass man jetzt sagen kann, in einem bestimmten Abstand kommt jetzt bestimmt wieder so ein blaues Dings-Bums da. **Bean:** Ich denke, das Modell dient nur dazu, das Ganze zu veranschaulichen, aber man kann anhand eines Modells keine Vorhersagen treffen.
0 - Antwortalternative wurde nicht entsprechend der theoretischen Intention verstanden	**Kaix:** Man erkennt im Modell zwar weitere [Bestandteile], die sind aber nicht beschriftet oder irgendwas. Und vorauszusagen war auch so ein bisschen komisch formuliert. - *Was findest du daran komisch?* - Naja vorauszusagen. Wenn, dann hätte ich jetzt anzugeben oder sowas eingesetzt, weil das ist irgendwie ein bisschen klarer dann.
8 - Antwort nicht aussagekräftig	**Kasa:** Ich habe mir davor noch die anderen [Antwortalternativen] durchgelesen und dachte, da passt eher so der Aufbau.

In der Aussage von Bean wird deutlich, dass es möglich ist, eine Antwortalternative auf dem theoretisch intendierten Niveau III zu verstehen, ohne selbst ein Modellverstehen auf Niveau III zu besitzen. Gleiches geschah bei Niveau II, wenn Schüler_innen anerkannten, dass man mit dem Modell auch erklären könnte, den Zweck dieses speziellen Modells aber nicht in einer Erklärung, sondern in der Beschreibung des Originals sahen. So sagt Gura: *Das erklärt nichts. Da ist keine Schrift, nichts beschrieben. Ich habe da eine Abbildung.*
Schüler_innen, die die Antwortalternative nicht entsprechend des Niveaus III verstanden, argumentierten hauptsächlich mit Bezug auf das Verb „voraussagen“, ohne dabei den hypothetischen Charakter der Aussage, welcher durch eben jenes Verb beschrieben werden

sollte, wiederzugeben. Sie interpretierten die Bedeutung des Verbs in eine Bedeutung um, welches den von ihnen gedachten Zweck des Modells besser beschreibt; in der Aussage von Kaix in das Verb „angeben" (Tab. 3).
Auch bei den angebotenen Antwortalternativen auf Niveau II, die nicht entsprechend der Theorie verstanden wurden, erkannten die Schüler_innen den Bedeutungsunterschied zwischen den Verben „zeigen/ darstellen/ wiedergeben" und „erklären/ erläutern/ verständlich machen" nicht und nutzen die Verben synonym. So begründete Sore seine völlige Zustimmung zur Niveau II Aussage „Das Modell der Biomembran hat den Zweck, den Aufbau der Biomembran zu erklären" wie folgt: *Die Bestandteile werden alle gezeigt und daran kann man veranschaulichen, wo die ganzen Proteine sich in der Membran befinden. Das kann man gut beschreiben.* Sore interpretierte die Antwortalternative in Niveau I um.
Zum Zweck der Selektion wurde für jede Antwortalternative eines jeden Kontexts ermittelt, wieviel Prozent der jeweils 40 interviewten Schüler_innen diese entsprechend des theoretisch intendierten Niveaus verstanden haben (Tab. 4).

Tabelle 4: Häufigkeit [%], mit der eine Antwortalternative entsprechend des theoretisch intendierten Niveaus interpretiert wurde. Graue Unterlegung: mehr als 70 %. Klammer: Anzahl der nicht interpretierbaren Aussagen von 40. $N_{Aussagen} = 1920$.

Kontext		Niveau I		Niveau II		Niveau III	
		Ia	Ib	IIa	IIb	IIIa	IIIb
Biomembran	[BM]	**98** (1)	93 (2)	**83** (2)	90 (4)	**75** (5)	53 (6)
Evolution	[EV]	75 (10)	**83** (4)	**74** (9)	85 (6)	24 (17)	**73** (7)
Gehirn	[GH]	**100** (0)	78 (9)	**73** (9)	44 (19)	**79** (4)	33 (9)
Jurawald	[JW]	**90** (4)	88 (5)	**70** (9)	53 (18)	38 (13)	**72** (7)
T. rex	[TR]	83 (7)	**83** (7)	69 (9)	**75** (10)	35 (13)	**70** (9)
Uferzone	[UF]	90 (4)	**95** (2)	**70** (10)	67 (12)	3 (9)	**70** (10)
Luftstrom	[LS]	63 (15)	58 (16)	53 (19)	53 (19)	63 (11)	45 (16)
Bakterienwachstum	[BW]	58 (14)	58 (16)	45 (16)	43 (18)	58 (13)	65 (12)

Legt man den für die Selektion definierten Richtwert an, dass mindestens 70 % der Schüler_innen eine Antwortalternative entsprechend des intendierten Niveaus verstanden haben müssen, dann findet man mit Ausnahme der Kontexte „Luftstrom" und „Bakterienwachstum" für jeden Kontext pro Niveau mindestens eine Antwortalternative, bei deren späterem Einsatz in Forced Choice Aufgaben ein zufriedenstellendes Verständnis der Schüler_innen erwartet werden kann.
Bei den Kontexten „Luftstrom" und „Bakterienwachstum" handelt es sich um Modelle, die als Formeln präsentiert werden. Beim „Luftstrom" wurde beispielsweise ein Schwimmer als Original dargestellt und eine Formel zur Menge der Luft in der Lunge beim Atmen. Viele Schüleraussagen waren für die Auswertung nicht aussagekräftig (Kodierung „8"; siehe Tab. 4), da die Schüler_innen die Formeln nicht als Modell verstanden. Der Proband Jaul sagt

doi: 10.4119/UNIBI/zdb-v1-i20-328

beispielsweise zu seiner Entscheidung in der Ratingaufgabe zum Kontext „Luftstrom“: *Da würde ich dann doch eher für das Modell hier [zeigt auf das Original] stimmen, weil da sieht man einen Menschen, der gerade ausatmet. Das [zeigt auf das Modell] ist hier eher wie eine Formel und ich finde da kann ich nicht so viel rauslesen, ehrlich gesagt, aus dieser Formel.* Alle Antwortalternativen zu den beiden Kontexten wurden im ersten Selektionsschritt verworfen.

Unterschiede zwischen theoretischer Intention und Interpretation durch die Schüler_innen zeigen sich sowohl zwischen den Niveaus als auch zwischen einzelnen niveaugleichen Antwortalternativen innerhalb eines Kontexts. Bei den sechs verbleibenden Kontexten nimmt bei höherem Niveau das theoriekonforme Verständnis der Antwortalternativen grundsätzlich ab. Auf Niveau III wird bei allen Kontexten jeweils nur eine der beiden Antwortalternativen von mind. 70 % der Schüler_innen entsprechend des Niveaus III interpretiert.

Im nächsten Schritt wurde pro Niveau die Relevanz der verstandenen Antwortalternativen untersucht. Für die Kontexte „Gehirn“, „Jurawald“, „T. rex“ und „Uferzone“ geschah dies ausschließlich auf Niveau I, für die Kontexte „Biomembran“ und „Evolution“ auf den Niveaus I und II.

5.2 Relevanz niveaugleicher Antwortalternativen

Vergleicht man die Bewertungen der Schüler_innen für die nach dem ersten Schritt verbleibenden Antwortalternativen pro Niveau pro Kontext auf der Likert-Skala, zeigen sich zum Teil deutliche Unterschiede (Abb. 2).

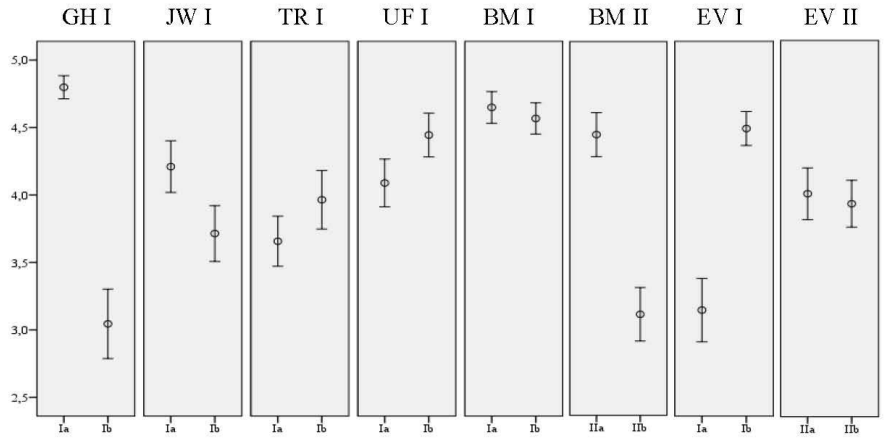


Abbildung 2: Bewertung der Antwortalternativen a und b pro Niveau [I, II, III] und pro Kontext (Abkürzungen [GH – EV] siehe Tab. 4.) auf der Likert-Skala (1: gar nicht, 2: wenig, 3: teils-teils, 4: annähernd, 5: völlig).

Beim Kontext der Biomembran (BM) wurde der Antwortalternative IIa auf der Likert-Skala deutlich mehr zugestimmt als der Antwortalternative IIb. Die Schüler_innen geben folglich an, dass das Modell der Biomembran eher den Zweck habe, den Aufbau der Biomembran zu erklären (Antwortalternative IIa), als das Zusammenwirken der Bestandteile begreiflich zu

machen (Antwortalternative IIb). In den mündlichen Begründungen argumentieren die Schüler_innen zwar, dass mit Hilfe des Modells auch einzelne Bestandteile erklärt werden können, das Zusammenwirken dieser mache das Modell aber durch das Fehlen schriftlicher Erklärungen oder der für das Verständnis von Prozessen notwendigen Animationen nicht begreiflich. Der Schüler Amon sagt im Interview dazu: *Das ist ein Bild und darauf bewegt sich nichts und dazu steht hier auch nichts. Demnach könnte ich mir so nicht erklären, wie die [Bestandteile] zusammenwirken.* Es wird deutlich, dass die Antwortalternative IIb von Amon entsprechend des theoretisch intendierten Niveaus II verstanden wurde, für ihn jedoch Mängel an diesem speziellen Modell den angegebenen Zweck für dieses Modell unterbinden. Beim Kontext Evolution (EV) stimmten die Schüler_innen den beiden Antwortalternativen auf Niveau II in etwa gleich stark zu. Antwortalternative IIa wurde selektiert, da sie sich auf den gleichen Aspekt des Kontexts bezieht, wie die Antwortalternativen aus den anderen Niveaus.
Für alle Forced Choice Aufgaben wurde aus den in Abbildung 2 dargestellten Antwortalternativen jeweils die auf der Likert-Skala höher bewertete und damit relevantere pro Kontext und Niveau ausgewählt. Die ausgewählten Antwortalternativen pro Kontext auf den Niveaus I-III sind in Tabelle 4 fett markiert.

6 Diskussion

Die Untersuchung beider Forschungsfragen trug dazu bei, insgesamt sechs Forced Choice Aufgaben zu verschiedenen biologischen Kontexten zu entwickeln. Die vorgestellte Prozedur schloss eine empirische Überprüfung der Eignung der Aufgaben für die Diagnose des Modellverstehens in der Teilkompetenz „Zweck von Modellen" durch Schüler_innen mit ein. Durch die Kombination eines quantitativen (Ratingskala) und eines qualitativen Ansatzes (Schülerinterviews) in einem *convergent mixed method design* (Creswell & Plano Clark, 2011) wurde eine Vielzahl von Schüleraussagen (N=1920) bei der Aufgabenentwicklung berücksichtigt. Dies kommt der Forderung nach, die Schüler_innen stärker in den Prozess der Aufgabenentwicklung einzubeziehen (Adams & Wieman, 2011; Leighton & Gierl, 2007).
Die Schüler_innen erklärten in Interviews, wie sie die einzelnen Antwortalternativen verstehen und lieferten dadurch bereits während der Aufgabenentwicklung Hinweise auf Validität. Die Ergebnisse dieses Vorgehens können jedoch nicht mit Protokollen lauten Denkens gleichgesetzt werden. Ericsson und Simon (1998) unterscheiden in Anlehnung an Vygotsky (1962) zwischen *inner speech* und *social speech* und betonen, dass die Verbalisierung der eigenen Gedankenprozesse in einer Gesprächssituation wie z. B. einem Interview (*social speech*) die Gedanken selbst ändern bzw. beeinflussen können. Kritiker schlussfolgern, dass es nicht möglich sei, die aufeinander folgenden Gedanken der Schüler_innen bei der Aufgabenbearbeitung zu erfassen, die zur Lösung führen (Ericsson & Simon, 1998). Diese Kritik kann für die vorliegende Studie dahingehend ausgeklammert werden, dass bei der Erhebung der Schüleraussagen nicht beabsichtigt wurde, den komplexen Gedankenprozess bei der Aufgabenbearbeitung nachzuvollziehen. Vielmehr sollte hier geprüft werden, ob die konstruierten Antwortalternativen wie intendiert verstanden wurden. Hierfür liefern die Interviewdaten interpretierbare und aussagekräftige Ansatzpunkte (Adams & Wieman, 2011).

doi: 10.4119/UNIBI/zdb-v1-i20-328

Bezogen auf die Ergebnisse der Untersuchung sind vor allem drei inhaltliche Aspekte zu diskutieren. Zum einen nimmt die theoriekonforme Interpretation der Antwortalternativen mit steigendem Niveau ab. Zum anderen scheinen die angebotenen Kontexte unterschiedliche kognitive Anforderungen zu transportieren, was teilweise zu Verständnisproblemen führt (Krell et al., 2014; Nehm & Ha, 2011). Es muss diskutiert werden, warum niveaugleiche Antwortalternativen zum gleichen Kontext für die Schüler_innen unterschiedlich relevant sind.

6.1 Einfluss des Niveaus auf die Interpretation der Antwortalternativen

Insgesamt zeigt sich eine gute Übereinstimmung zwischen dem theoretisch intendierten Niveau in den Antwortalternativen und dem Verstehen dieser durch die Schüler_innen (Forschungsfrage 1). Folglich können viele der konstruierten Antwortalternativen dazu dienen, Rückschlüsse auf das Verstehen der Schüler_innen zu ziehen (Tab. 4; AERA et al., 2014). Nichtsdestotrotz nimmt die Passung des interpretierten mit dem theoretisch intendierten Niveau der Antwortalternativen mit höheren Niveaus ab. Als ein Erklärungsansatz könnte ein geringes Modellverstehen von Seiten der Schüler_innen herangezogen werden. Geht man davon aus, dass Schüler_innen die Rolle von Modellen im wissenschaftlichen Erkenntnisprozess in der Schule kaum erleben (Crawford & Cullin, 2005; Justi & Gilbert, 2005; van Driel & Verloop, 2002) und Modelle daher nicht als Instrumente der Forschung ansehen (Grosslight et al., 1991; Grünkorn, 2014; Treagust et al., 2002; Trier & Upmeier zu Belzen, 2009), ist es möglich, dass sie dem Inhalt von Antwortalternativen, die die epistemologische Bedeutung von Modellen ausdrücken (Niveau III), nicht nur nicht zustimmen, sondern diesen Inhalt auch nicht entsprechend verstehen. Als Folge zeigt sich möglichweise, dass das niveauangebende Verb (z. B. vorhersagen, vermuten) von Schüler_innen ignoriert und die Zustimmung zu einer Antwortalternative allein nach deren Inhalt entschieden wird. Alternativ könnte eine Uminterpretation des niveauangebenden Verbs entsprechend des eigenen Modellverstehens erfolgen. Diese Art der Uminterpretation als Strategie wurde auch von Krell, Czeskleba und Krüger (2012) in einer Studie mit lautem Denken identifiziert. Die Forscher beobachteten, dass Schüler_innen in Paarvergleichs-aufgaben, sofern ihr präferiertes Niveau nicht angeboten wurde, eine andere Antwortalternative entsprechend des gewünschten Niveaus umdeuteten. Die valide Interpretation der Ergebnisse bezogen auf das Modellverstehen ist in diesen Fällen nicht möglich (AERA et al., 2014).

Im Hinblick auf die Nutzung der Forced Choice Aufgaben zur Diagnose von Modellverstehen führen Uminterpretationen oder Bewertungen auf der Basis des Inhalts der Antwortalternativen dazu, dass Schüler_innen ein nicht zutreffendes Modellverstehen zugeschrieben wird. Dies führt wiederum zu Problemen bei der adäquaten Förderung. Die Interviewdaten konnten hier genutzt werden, um Hinweise für die Auswahl von Antwortalternativen zu sammeln und damit diese Probleme zu reduzieren.

doi: 10.4119/UNIBI/zdb-v1-i20-328

6.2 Einfluss des Kontexts auf die Interpretation der Antwortalternativen

Für die Kontexte „Luftstrom“ und „Bakterienwachstum“ war es nicht möglich, Antwortalternativen zu selektieren, die von den Schüler_innen in ausreichendem Maße wie intendiert verstanden wurden. Hier konnte die Hypothese, dass die Bearbeitung der Aufgaben später zu valide interpretierbaren Diagnoseergebnissen führt, nicht bestätigt werden (AERA et al., 2014). Gründe für das Nichtverstehen der Antwortalternativen scheinen nicht niveauspezifisch zu sein, da der Prozentsatz der Schüler_innen, die die Antwortalternativen nicht verstehen, sich zwischen den Niveaus I, II und III nicht unterscheidet. Es kann vermutet werden, dass die Repräsentationsform der Modelle (hier Formeln) ein Grund für die Verständnisprobleme war. Cohors-Fresenborg et al. (2004), die PISA Aufgaben im Hinblick auf kognitionsorientierte Aufgabenmerkmale untersuchten und klassifizierten, leiten aus ihren Ergebnissen ab, dass das Merkmal „Formalisierung von Wissen“ ein auf besondere Weise schwierigkeitserzeugendes Merkmal ist. Cohors-Fresenborg et al. (2004) betonen zudem, es handle sich bei formalisiertem Wissen auch um ein „Werkzeug, dessen Handhabung eine Kompetenz darstellt, Komplexität zu bewältigen“ (Cohors-Fresenborg et al., 2004, S. 121). Möglicherweise können die Schüler_innen mit diesem Werkzeug noch nicht entsprechend kompetent umgehen. Dies zeigt sich womöglich in der Schüleraussage von Jaul: *Weil das ja hier eher wie so eine Formel ist und ich finde, da kann ich nicht so viel rauslesen, ehrlich gesagt, aus dieser Formel.* Solche und ähnliche Aussagen lassen die Vermutung zu, dass die Aufgaben mit den formelhaften Modellen zu den Kontexten „Luftstrom“ und „Bakterienwachstum“ nicht nur das Modellverstehen der Schüler_innen, sondern, als eine Interaktion, möglicherweise auch deren Umgang mit Formeln messen. Durch die mangelnde Trennung der Konstrukte eignen sich diese Kontexte nicht zur Erfassung von Modellverstehen und wurden daher verworfen.

6.3 Unterschiede zwischen niveaugleichen Antwortalternativen zu einem Kontext

Scheinbar führen kontextspezifische inhaltliche Aspekte dazu, dass Schüler_innen eine von zwei niveaugleichen Antwortalternativen mehr oder weniger bevorzugen. Die Gründe könnten bei den inhaltlichen Eigenschaften der Antwortalternativen liegen. Die Inhalte einiger Antwortalternativen passen nach Meinung der Schüler_innen offensichtlich besser oder weniger gut zum dargestellten Modellobjekt als andere. Demnach wird bei einigen Antwortalternativen eine andere Repräsentation des Modells erwartet, z. B. eine Beschriftung oder Animation der Biomembran.

Außerdem spielen personenbezogene Merkmale wie Vorwissen und Interesse eine Rolle. Obwohl das Vorwissen bei dieser Studie nicht erhoben wurde, zeigen sich in den Interviewdaten einige interessante Schülervorstellungen, die sich möglicherweise auf die Bewertung der Antwortalternativen auswirken. Beim Kontext des Bakterienwachstums zeigt sich die Schülervorstellung, dass Lebewesen durch Vergrößerung der Zellen wachsen („Wachstum ist größer werden“; Riemeier, 2005) z. B. darin, dass viele Schüler_innen den Ausdruck „Menge an Bakterien“ kritisch bewerten. Douc begründet: *Da habe ich wenig angekreuzt, weil es um das Wachstum geht und nicht um die Menge an Bakterien.*

doi: 10.4119/UNIBI/zdb-v1-i20-328

Mit Bezug zum eigenen Interesse kommentiert Jael die Bevorzugung der Antwortalternative „Das Modell des T. rex hat den Zweck, den Einfluss der Vorderbeine auf die Fortbewegungsart des T. rex verständlich zu machen." gegenüber der niveaugleichen Antwortalternative wie folgt: *[Es ist] ziemlich interessant, was die Vorderbeine eigentlich für einen Einfluss haben darauf, wie er sich bewegt.* Solche inhaltlichen Aspekte werden von Kauertz (2008) als potentiell schwierigkeitserzeugend beschrieben. Möglicherweise wirkt sich das Interesse nicht nur darauf aus, welche Antwortalternative als besonders relevant in Bezug auf den Kontext wahrgenommen wird, sondern auch darauf, welchen Zweck das Modell für eine Schüler_in besitzt. Werner, Schwanewedel und Mayer (2014) zeigen, dass für ihre Untersuchung unterschiedlicher Kontexte bei der Bewertungskompetenz alle Kontext-Personen-Valenzen (z. B. Bekanntheit, Alltags- und Gesellschaftsrelevanz, Interessantheit) positiv mit der Personenfähigkeit korrelierten. In Anlehnung daran kann argumentiert werden, dass der fokussierte Aspekt des biologischen Kontexts nicht in sich selbst leicht oder schwer für Schüler_innen ist, sondern bekannt, relevant oder interessant sein kann und daher von Schüler_innen unterschiedlich bewertet wird. Hieraus lässt sich für die Zusammenstellung der Forced Choice Aufgaben ableiten, dass möglichst Antwortalternativen zusammen präsentiert werden sollten, die sich auf den gleichen Aspekt des Kontexts beziehen, da sonst eine Entscheidung rein nach inhaltlichem Interesse nicht ausgeschlossen werden kann und eine valide Interpretation der Diagnoseergebnisse unmöglich wird (AERA et al., 2014).

7 Fazit und Ausblick

Bei der Entwicklung von Diagnoseaufgaben für Schüler_innen fordern aktuelle Standards der Testentwicklung, die Schüler_innen selbst als Zielstichprobe mehr einzubeziehen (AERA et al., 2014; NRC 2011; Leighton, 2004). Dabei gilt es für eine valide Interpretation der Ergebnisse abzusichern, dass die Schüler_innen das mit dem Test zu messende Wissen und Können auch tatsächlich bei der Bearbeitung der Aufgaben nutzen (AERA et al., 2014). Das in diesem Artikel vorgestellte Vorgehen zur Einbeziehung von Schüler_innen in den Prozess der Aufgabenentwicklung am Beispiel von Diagnoseaufgaben zum Zweck von Modellen machte es möglich, Schwierigkeiten in zuvor konstruierten Antwortalternativen offen zu legen. Die Verständnisprobleme auf Seiten der Schüler_innen würden eine valide zu interpretierende Diagnose unterbinden. Auf der Grundlage der Schülerbewertungen in den Ratingaufgaben und den Schüleraussagen in den Interviews konnten begründet Antwortalternativen selektiert und in Forced Choice Aufgaben zusammengestellt werden.

In einem nächsten Schritt gilt es, die zusammengestellten Forced Choice Aufgaben bei Schüler_innen einzusetzen und empirisch zu überprüfen, ob diese Aufgaben die Hypothese einer validen Diagnose weiter unterstützen. Hierfür werden die Eigenschaften der Aufgaben in einer Studie mittels einer Kombination aus Eyetracking und retrospektivem lautem Denken untersucht (Antwortprozesse; AERA et al., 2014). Eine Dimensionalitätsanalyse wird Aufschlüsse geben, inwieweit die in den Daten ersichtlichen Zusammenhänge zwischen den Aufgaben die theoretisch erwarteten Zusammenhänge bestätigen (Interne Struktur; AERA et al., 2014). Der Vergleich der Diagnoseergebnisse mit externen Variablen und die Überprüfung der Sensibilität der Diagnose bezogen auf Fördermaßnahmen wird die Validität

doi: 10.4119/UNIBI/zdb-v1-i20-328

der Ergebnisinterpretation weiter hinterfragen (Beziehungen zu anderen Variablen und Außenkriterien; AERA et al., 2014).

Literatur

Adams, W. K. & Wieman, C. E. (2011). Development and Validation of Instruments to Measure Learning of Expert-Like Thinking. *International Journal of Science Education, 33* (9), 1289-1312.

AERA, APA & NCME [American Educational Research Association, American Psychological Association & National Council on Measurement in Education] (Hrsg.). (2014). *Standards for educational and psychological testing*. Washington, DC: American Educational Research Association.

Cohors-Fresenborg, E., Sjuts, J. & Sommer, N. (2004). Komplexität von Denkvorgängen und Formalisierung von Wissen. In M. Neubrand (Hrsg.), *Mathematische Kompetenzen von Schülerinnen und Schülern in Deutschland. Vertiefende Analysen im Rahmen von PISA 2000* (S. 109-144). Wiesbaden: VS Verlag für Sozialwissenschaften.

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. Eijkelhoff (Hrsg.), *Research and the quality of science education* (S. 309-323). Dordrecht: Springer.

Creswell, J. & Plano Clark, V. (2011). *Designing and conducting mixed methods research*. Los Angeles, CA: Sage.

Ericsson, K. A. & Simon, H. A. (1998). How to study thinking in everyday life: Contrasting think-aloud protocols with descriptions and explanations of thinking. *Mind, Culture, and Activity, 5* (3), 178-186.

Fleischer, J., Koeppen, K., Kenk, M., Klieme, E. & Leutner, D. (2013). Kompetenzmodellierung: Struktur, Konzepte und Forschungszugänge des DFG-Schwerpunktprogramms. *Zeitschrift für Erziehungswissenschaft, 16* (1), 5-22.

Gonzalez, E. & Rutkowski, L. (2010). Principles of multiple matrix booklet designs and parameter recovery in large-scale assessments. *IERI monograph series: Issues and methodologies in large-scale assessments, 3,* 125-156.

Gorin, J. S. (2007). Test Construction and Diagnostic Testing. In J. Leighton & M. Gierl (Hrsg.), *Cognitive Diagnostic Assessment for Education. Theory and Applications* (S. 173-202). Cambridge: Cambridge University Press.

Grosslight, L., Jay, E., Unger, C. & Smith. (1991). Understanding models and their use in science. Conceptions of middle and high school students and experts. *Journal of Research in Science Teaching, 28* (9), 799-822.

Grünkorn, J. (2014). *Modellkompetenz im Biologieunterricht. Empirische Analyse von Modellkompetenz bei Schülerinnen und Schülern der Sekundarstufe I mit Aufgaben im offenen Antwortformat.* Dissertation.

Haladyna, T. M. (2004). *Developing and validating multiple-choice test items* (3rd ed.). Mahwah, NJ: Lawrence Erlbaum Associates.

Harrison, A. G. & Treagust, D. F. (2000). A typology of school science models. *International Journal of Science Education, 22* (9), 1011-1026.

Hartig, J., Klieme, E. & Leutner, D. (2008). *Assessment of competencies in educational contexts: State of the art and future prospects.* Göttingen: Hogrefe & Huber.

Hodson, D. (2014). Learning Science, Learning about Science, Doing Science. Different goals demand different learning methods. *International Journal of Science Education, 36* (15), 2534-2553.

Ingenkamp, K. & Lissmann, U. (2008). *Lehrbuch der pädagogischen Diagnostik.* Weinheim: Beltz.

Justi, R. S. & Gilbert, J. K. (2005). Investigating teachers' ideas about models and modelling: some issues of authenticity. In K. Boersma, M. Goedhart, O. de Jong & H. Eijkelhoff (Hrsg.), *Research and the quality of science education* (S. 325-335). Dordrecht: Springer.

Kauertz, A. (2008). *Schwierigkeitserzeugende Merkmale physikalischer Leistungsaufgaben.* Berlin: Logos.

Kauertz, A., Neumann, K. & Haertig, H. (2012). Competence in Science Education. In B. J. Fraser, K. Tobin & C. J. McRobbie (Hrsg.), *Second International Handbook of Science Education* (S. 711-721). Dordrecht: Springer.

doi: 10.4119/UNIBI/zdb-v1-i20-328

Klieme, E. & Hartig, J. (2007). Kompetenzkonzepte in den Sozialwissenschaften und im erziehungswissenschaftlichen Diskurs. *Zeitschrift für Erziehungswissenschaft, 10* (8), 11-29.

KMK [Ständige Konferenz der Kultusminister der Länder in der Bundesrepublik Deutschland]. (2005). Standards für die Lehrerbildung: Bildungswissenschaften.: Beschluss der Kultusministerkonferenz vom 16.12.2004. *Zeitschrift für Pädagogik, 51* (2), 280-290.

Krell, M., Czeskleba, A. & Krüger, D. (2012). Validierung von Forced Choice-Aufgaben durch Lautes Denken, *Erkenntnisweg Biologiedidaktik, 11*, 53-70.

Krell, M. & Krüger, D. (2013). Wie werden Modelle im Biologieunterricht eingesetzt? Ergebnisse einer Fragebogenstudie, *Erkenntnisweg Biologiedidaktik,12,* 9-26.

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. Verfügbar unter http://www.ijobed.com/2_2/Moritz-2012.pdf

Krell, M., Upmeier zu Belzen, A. & Krüger, D. (2014). 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 (Hrsg.), *E-Book proceedings of the ESERA 2013 conference. Science education research for evidence-based teaching and coherence in learning. Part 6. Nature of science: History, philosophy and sociology of science.* Nicosia, Cyprus: European Science Education Research Association.

Leighton, J. P. & Gierl, M. (2007). Defining and evaluating models of cognition used in educational measurement to make inferences about examinees' thinking processes. *Educational Measurement: Issues and Practice, 26* (2), 3-16.

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* (4), 6-15.

Mahr, B. (2008). Ein Modell des Modellseins. Ein Beitrag zur Aufklärung des Modellbegriffs. In U. Dirks & E. Knobloch (Hrsg.), *Modelle* (S. 187-218). Frankfurt am Main: Peter Lang.

McCloy, R., Heggestad, E., & Reeve, C. (2005). A silk purse from the sow's ear: retrieving normative information from multidimensional forced-choice items. *Organizational Research Methods, 8*, 222-248.

Moosbrugger, H. & Kelava, A. (2012). *Testtheorie und Fragebogenkonstruktion*. Berlin: Springer.

NRC [National Research Council]. (2001). *Knowing What Students Know*. Washington, DC: National Academies Press.

Nehm, R. H. & Ha, M. (2011). Item feature effects in evolution assessment. *Journal of Research in Science Teaching, 48* (3), 237-256.

Passmore, C., Gouvea, J. & Giere, R. (2014). Models in science and in learning science. In M. Matthews (Hrsg.), *International handbook of research in history, philosophy and science teaching* (S. 1171-1202). Dordrecht: Springer.

Riemeier, T. (2005). Schülervorstellungen von Zellen, Teilung und Wachstum. *Zeitschrift für Didaktik der Naturwissenschaften, 11* (1), 52-72.

Sandmann, A. (2014). Lautes Denken - die Analyse von Denk-, Lern- und Problemlöseprozesse. In D. Krüger, I. Parchmann & H. Schecker (Hrsg.), *Methoden in der naturwissenschaftsdidaktischen Forschung* (S. 179-188). Heidelberg: Springer.

Treagust, D. D., Chittleborough, G. D. & Mamiala, T. L. (2002). Students' understanding of the role of scientific models in learning science. *Journal of Science Education, 24* (4), 357-368.

Trier, U. & Upmeier zu Belzen, A. (2009). „Die Wissenschaftler nutzen Modelle, um etwas Neues zu entdecken, und in der Schule lernt man einfach nur, dass es so ist." Schülervorstellungen zu Modellen. *Erkenntnisweg Biologiedidaktik, 8,* 23-37.

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

Van Driel, J. H. & Verloop, N. (2002). Experienced teachers' knowledge of teaching and learning of models and modelling in science education. *International Journal of Science Education, 24* (12), 1255-1272.

Vygotsky, L. (1962). *Thought and language*. Cambridge, MA: MIT Press.

doi: 10.4119/UNIBI/zdb-v1-i20-328

Werner, M., Schwanewedel, J. & Mayer, J. (2014). Does the context make a difference? Students' abilities in decision-making and the influence of contexts. In C. Constantinou, N. Papadouris & A. Hadjigeorgiou (Hrsg.), *E-Book proceedings of the ESERA 2013 conference. Science Education Research For Evidence-based Teaching and Coherence in Learning. Part 8. Scientific Literacy and socio scientific issues* (S. 81-89). Nicosia, Cyprus: European Science Education Research Association.

Wirtz, M. & Caspar, F. (2002). *Beurteilerübereinstimmung und Beurteilerreliabilität. Methoden zur Bestimmung und Verbesserung der Zuverlässigkeit von Einschätzungen mittels Kategoriensystemen und Ratingskalen.* Göttingen: Hogrefe.

Kontakt

Sarah Gogolin
Didaktik der Biologie
Schwendenerstraße 1
14195 Berlin
sarah.gogolin@fu-berlin.de

doi: 10.4119/UNIBI/zdb-v1-i20-328

Article 4 Gogolin, S. & Krüger, D. (submitted). Students' understanding of the nature and purpose of models. *Journal of Research in Science Teaching*.

[Manuscript as archivable pre-print]

Sarah Gogolin & Dirk Krüger

Students' understanding of the nature and purpose of models

Abstract

The process of thinking in and about models as a scientific practice should be integrated into science teaching and learning. Empirical studies show that students see models primarily in their role as media to facilitate content learning whilst rarely appreciating models as instruments of scientists which allow the deduction and the testing of predictions. In order to foster their students successfully, teachers need specific diagnostic information about their students' understanding of models and modelling. The aim of this study is to provide teachers with this information by investigating students' understanding of the nature and the purpose of models in biology with respect to context- and grade-specific differences. Students' understanding of the nature and the purpose of models was assessed by using forced choice tasks ($N_{students} = 285$). In order to gain qualitative insight into possible reasons for context-specific differences, the students were additionally asked to give reason for their decision in the forced choice tasks by answering open-ended justification tasks. The results indicate that the majority of students in all grades see models as idealized representations of an original which have the purpose to show or to describe this original. Students' levels of understanding of the nature and the purpose of models increase only little across grades. Nevertheless, a grade-specific analysis of the consistency of students' understanding across contexts suggests that the students' understanding becomes more consistent in higher grades. Students' justifications helped to identify model contexts which have a high potential to be fruitful when trying to foster students' understanding of the nature and the purpose of models.

Keywords: models, biology education, students' understanding, diagnosis, contexts

Introduction

In an open letter to the World Health Organization, a group of international scientists requested for the Olympic Games 2016 to be postponed "in the name of public health". On the basis of their models, the scientists predicted a dangerous worldwide spread of the Zika-Virus by the Games. The operations behind this example address the five pragmatic uses for models in biology which were outlined by philosopher of science Jay Odenbaugh (2005): Scientists use models to (1) explore complex systems, (2) explore unknown possibilities (3) develop conceptual frameworks, (4) make accurate predictions and (5) generate causal explanations. The use of models in these ways should not be reserved for scientists alone, but should be integrated into science teaching and learning (e.g., Gobert et al., 2011; Passmore, Gouvea, & Giere, 2014; Upmeier zu Belzen & Krüger, 2010). Following Hodson (2014), students may use models in order to gain conceptual and theoretical knowledge ('learn science'), to engage in scientific practices ('do science') and to develop an understanding of the characteristics of science as part of nature of science ('learn about science'). The process of thinking in and about models as one of eight scientific practices specially highlighted in NGSS (NGSS Lead States, 2013) can be seen as one of the essential learning goals for science students in order to understand nature of science and develop scientific literacy. Empirical studies with students, however, indicate that the latter see models primarily in their role as media to describe complex phenomena and to facilitate content learning whilst rarely appreciating models as instruments of scientists which allow the deduction and the testing of predictions (Chittleborough, Treagust, Mamiala, & Mocerino, 2005; Grosslight, Jay, Unger, & Smith, 1991; Grünkorn, Upmeier zu Belzen, & Krüger, 2014; Krell, Upmeier zu Belzen, & Krüger, 2014c; Treagust, Chittleborough, & Mamiala, 2002). Consequently, teachers need to engage their students to 'learn about science' by reflecting about this second perspective on models. Science education researchers ask for science teacher education courses to cultivate teachers' understanding of models and modeling and equip future teachers with relevant skills and diagnostic information about their students' understanding (e.g., Günther, Fleige, Upmeier zu Belzen, & Krüger, 2016; Henze, van Driel, & Verloop, 2008; Justi & van Driel, 2005, 2006; Vo et al., 2015). It is necessary for teachers to know about the status quo of their students' understanding in order to customize lesson planning and teaching (Campbell, Schwarz, & Windschitl, 2016; Duit, Gropengießer, Kattmann, Komorek, & Parchmann, 2012). We believe that the report of diagnostic information about students' understanding of models and modeling should be as specific as possible and take into consideration variations across

contexts and age. This study aims to investigate students' understanding of the nature and the purpose of models in biology. We present a closed-ended instrument that diagnoses individual students' understanding of models and modeling with respect to context- and grade-specific differences on the basis of a theoretical framework that conceptualizes students' understanding of the nature and purpose of models. The results of the diagnosis will be used to make propositions for adequate ways of teaching about models and modeling.

Theoretical Background

What is a model? – Giving a straightforward answer to this question is very challenging due to the omnipresence and versatility of models in both everyday life and in science (cf. Oh & Oh, 2011). Consequently, we will try to step away from a definition for the term 'model' and instead refer to Mahr (2008, 2011) who determines interdependent relationships that justify something to be conceived of as a model. In his epistemic pattern of model-being, a model is (1) distinct from its representation as a model-object, (2) in its function as a representation, a model *of* something, and (3), in a methodological view, a model *for* something (Mahr, 2008, 2011). For example, a computer simulation of the Zika-Virus' spread (the model-object) is based on assumptions drawn from recorded data with regard to the rate of infection transmission, thus being a model *of* the original Zika-Virus' spread. Moreover, as a model *for* predicting, it allows the testing of hypotheses, which are drawn from the model itself about the number of lethal infections or about the potential spatial dispersion of the Zika-Virus by the Olympic Games. Mahr (2011) points out that models "do not incarnate any form of truth, but rather forms of demonstrability, possibility, and choice" (Mahr 2011, p. 303). The distinction between 'model *of*' and 'model *for*' can also be used to describe the purposes of models as do Cartier, Rudolph, and Stewart (2001) who argue that "scientific models are both desirable products of scientific research and useful as guides to future research" (p. 2). Summarizing purposes attributed to models by science philosophers and science education researchers, Oh and Oh (2011) highlight that "as description, explanation and prediction are primary goals of science, the purposes of modeling in science are to describe, explain and predict particular aspects of the natural world" (pp. 1114-1115). Despite the many divergent attempts to define models and their use in science, researchers agree that models are the result of a theory-driven modeling process and are defined only in the context of their use (e.g.,

Bailer-Jones, 2002; Giere, 2001; Mahr, 2008, 2011; Odenbaugh, 2005; van der Valk, van Driel, & Vos, 2007).

What do we want our students to learn with regard to models?

Among science education researchers there is an agreement that models and modeling should be part of the science curriculum (e.g., Campbell, Oh, Maughn, Kiriazis, & Zuwallack, 2015; Gilbert & Justi, 2016; Oh & Oh, 2011; Passmore et al., 2014). At the same time, there is a multitude of pedagogical aims that may be reached by employing models. In their review article about modeling pedagogies, Campbell et al. (2015) report that "conceptual understanding was the most common pedagogical function identified for modeling, while developing facility and understanding of science practices was identified least often" (p. 159). Referring to Hodson's (2014) terminology, this means that models are more often being used for 'learning science' than for 'learning about science'. Science education standard documents across countries (e.g., Germany: KMK, 2005; UK: QCA, 2007; USA: NGSS Lead States, 2013) call for the integration of and the reflection about modeling as a science practice into science lessons. Schwarz et al. (2009) ask for students to develop a meta-modeling knowledge which enables them to reflect about "how models are used, why they are used, and what their strengths and limitations are, in order to appreciate how science works and the dynamic nature of knowledge that science produces" (pp. 634–635). Such 'epistemic considerations' are distinct from modeling practices (cf. Vo et al., 2015). The combination of both dimensions (meta-modeling knowledge and modeling practices) forms model competence (Gilbert & Justi, 2016; Nicolaou & Constantinou, 2014; Upmeier zu Belzen & Krüger, 2010).

A variety of teachers' and/or students' perspectives with regard to models and modeling are conceptualized in different theoretical frameworks (Crawford & Cullin, 2005; Grosslight et al., 1991; Justi & Gilbert, 2003; Schwarz et al., 2009; Treagust et al., 2002; Upmeier zu Belzen & Krüger, 2010). The theoretical framework for the present article is the 'model of model competence' (Upmeier zu Belzen & Krüger, 2010; cf. Grünkorn et al., 2014) which describes the five aspects 'nature of models', 'alternative models', 'purpose of models', 'testing models' and 'changing models'. For each of the aspects, the authors propose three levels which reflect increasingly sophisticated ways of understanding the aspect. The two aspects 'nature of models' and 'purpose of models' which form the theoretical basis for the present article are presented in more detail in Table 1.

Table 1
Theoretical framework of model competence (Grünkorn et al., 2014; Upmeier zu Belzen & Krüger, 2010)

	Level I	Level II	Level III
Nature of models	Model is a replication of the original	Model is an idealised representation of the original	Model is a theoretical reconstruction of the original
Purpose of models	Using the model to describe the original	Using the model to explain something about the original	Using the model to predict something about the original

According to Upmeier zu Belzen and Krüger (2010), a scientific understanding of the nature of models and/or the purpose of models is accredited to students who are able to understand all perspectives in the particular aspect.

Empirical findings on perspectives towards models

Empirical studies indicate that teachers' (Borrmann, Reinhardt, Krell, & Krüger, 2014; Crawford & Cullin, 2005; Danusso, Testa, & Vicentini, 2010; Justi & Gilbert, 2002, 2003; Krell & Krüger, 2016; van Driel & Verloop, 1999, 2002) and students' (Chittleborough et al., 2005; Gobert et al., 2011; Grosslight et al., 1991; Grünkorn et al., 2014; Krell et al., 2012, 2014a, 2014c; Lee et al., 2015; Patzke et al., 2015; Treagust et al., 2002, 2004; Trier, Krüger, & Upmeier zu Belzen, 2014) knowledge of and about models is rather limited. Students' understanding seems to be bound primarily to the view on models as representations of concrete objects that ought to be as similar as possible to the original (Grosslight et al., 1991; Grünkorn et al., 2014). Concerning the purpose of models, students tend to differentiate between the school and the science context (Trier et al., 2014; N=7) but few students are aware of the modeler as the agent that determines the purpose of the model (Grosslight et al., 1991; N=55). In a study with open-ended diagnostic tasks, Grünkorn et al. (2014; N=706) found that all of the students believed models in biology to serve for showing facts or for identifying or explaining relationships in order to understand these facts. 24 % of the students additionally stated that models were instruments to examine ideas by testing hypotheses about the original. Interview studies report that students do not appreciate models to be constructed in the service of developing and testing ideas (Grosslight et al., 1991; Trier et al., 2014). Some studies show that students' meta-modeling knowledge and modeling practices can be

fostered successfully by modeling-based instruction (Capps, Shemwell, Lindsey, Gagnon, & Owen, 2016; Gilbert & Justi, 2016; Hergert, Krell, & Krüger, in prep; Sins, Savelsbergh, van Joolingen, & van Hout-Wolters, 2009; Zangori, Forbes, & Schwarz, 2015).

The previous, rather general description of students' perspectives regarding the nature and the purpose of models has to be treated with caution as there are a number of studies questioning a general approach to assessing students' understanding of models and modeling (e.g., Chittleborough et al., 2005; Krell et al., 2012). These studies investigated whether students' understanding of models is context- and/or grade-specific.

Students' understanding across contexts

Science education standards point out that engaging in modeling is only really an epistemic practice of science in the presence of real disciplinary phenomena (NGSS Lead States, 2013). Passmore et al. (2014) conclude from the condition that models are defined only in the context of their use that "there is no context-independent way to evaluate a model. Models are built with an understanding of the epistemological criteria that are relevant to the question at hand, and they are evaluated with an understanding of their intended use" (p. 1183).

The term 'context' is being used broadly in the field of science education, covering (1) contexts as task features (e.g., employing different models in task stems; Al-Balushi, 2011, Krell et al., 2012), (2) contexts as discipline-specifications (e.g., tasks in biology, chemistry, physics; Gobert et al., 2011, Krell, Reinisch, & Krüger, 2015) and (3) contexts as learning situations (e.g., support in learning; Berland & Crucet, 2016, van Oers, 1998). In the present study, the term 'context' refers to different models as features in the task stems.

It has widely been reported that students' understanding of models seems to be context-specific (e.g., Al-Balushi, 2011; Krell et al., 2012, 2014a; Krell & Krüger, in prep.; Lee et al., 2015; Pluta et al., 2011). Krell et al. (2012) assessed students' (N=1,207) understanding of the purpose of models using one decontextualized as well as six contextualized forced choice tasks and documented inconsistencies across different biological contexts. Krell et al. (2014a) point out that the form of epistemic knowledge about models (e.g., context-specific or -independent) is an important issue for science education because it has consequences for both assessment and teaching. While contextualized tasks are being discussed as an issue of assessment by quite a number of researchers who refer to models and modeling (e.g., Al-Balushi, 2011; Grünkorn et al., 2014; Krell & Krüger, in press; Krell et al., 2012, 2014a; Lee et al., 2015) as well as in other domains of science education research (e.g., Guerra-Ramos,

2012; Leach, Millar, Ryder, & Séré, 2000; Nehm & Ha, 2011), the issue of contextualized teaching with regard to models and modeling has not been focused on as much (Krell, 2013). The importance of the context for learning has been pointed out by developmental psychology researchers who see context as part of behavior (e.g., Fischer, Bullock, Rotenberg, & Raya, 1993; Kitchener, Lynch, Fischer, & Wood, 1993; van Geert, 1998; van Geert & van Dijk, 2002). Fischer et al. (1993) highlight that „Context does not merely influence behavior. It is literally part of the behavior, participating with the person to produce an action or thought." (p. 93). According to Fischer et al. (1993), traditional conceptions that treat competence as a fixed characteristic of the child fail because children show different levels of competence in different contexts. In so called 'optimal contexts', children would show a true upper limit on performance, called optimal level. In 'spontaneous contexts', on the other side, children would show a much lower upper limit, called functional level (Fischer et al., 1993). Following this distinction, we have to assume some model contexts to be rather 'optimal' and other to be rather 'spontaneous' for students. Engaging students with optimal contexts may contribute to their learning as a form of scaffolding (Fischer et al., 1993). Passmore et al. (2014) argue that teachers should carefully choose which canon of models to integrate into their science classes. Krell (2013) has applied the distinction between 'optimal' and 'spontaneous' contexts to understanding models and concludes that there is too little evidence as to which characteristics 'optimal' and 'spontaneous' contexts may have. More investigations into how exactly students understand models across different contexts is needed to identify optimal model contexts which have a high potential to be fruitful when trying to foster students' understanding of the nature and the purpose of models.

Students' understanding across grades

Several studies which included subgroups with students of different grades into their investigations report that the age of students has an influence on their understanding of models, with older students answering more elaborately (e.g., Al-Balushi, 2011; Chittleborough et al., 2005; Grosslight et al., 1991; Lee et al., 2015). With respect to the aspect 'nature of models', Al-Balushi (2011) points out that there is a "drop in the number of students across grade levels who believe that there is a 1:1 correspondence between models and reality" (p. 575). Regarding the aspect 'purpose of models', Chittleborough et al. (2005) state: "With increasing age and maturity, more students per year, from grade 8 to first year of university, were able to describe the role of the scientific model in the scientific process"

(p. 210). Other findings do not confirm the aforementioned differences (Grünkorn et al., 2014; Krell et al., 2014c; Patzke et al., 2015; Treagust et al., 2002). The SUMS instrument of Treagust et al. (2002), showed "no statistically significant differences for any of the scales between year levels" (p. 361) for the grades eight to ten. Patzke et al. (2015) conducted a longitudinal study over the count of three years (beginning with seventh grade) and report merely very small changes in the students' understanding of models and modeling. Some researchers draw a connection between contexts and developmental stage and conclude that not necessarily the level but the consistency of students' understanding across contexts is positively correlated with the age of the students (e.g., Clough & Driver, 1986; Fischer et al., 1993; Kitchener et al., 1993; Krell et al., 2014c).

Studies investigating students' understanding of models and modeling across grades employed methods that permitted to detect to what extent the subgroups differ but they do not describe the qualities of the differences in terms of distinct levels of understanding models and modeling (e.g., Chittleborough et al., 2005; Lee et al., 2015; Patzke et al., 2015; Treagust et al., 2002). A comprehensive and easily interpretable diagnosis that reports the dispersion of students' levels of understanding of the nature and the purpose of models in different grades has more chance to help teachers plan appropriate promotions.

Research questions

Oh and Oh (2011) argue that it is indispensable for teachers of science to be aware of the notions of models for they can use models effectively in their science classrooms. Günther et al. (2016) specify that this knowledge alone does not suffice to foster students' understanding successfully. Differentiated diagnostic information is a prerequisite for science teachers who are wishing to foster their students' understanding of models and modeling (Campbell et al., 2016; Duit et al., 2012; Henze et al., 2008; Justi & van Driel, 2005, 2006). We aim at investigating into students' understanding of the nature and the purpose of models in biology with respect to context- and grade-specific differences by using a closed-ended instrument that diagnoses individual students' meta-knowledge about the nature and the purpose of models. In order to supply teachers with information about students' understanding of models across contexts and across grades, we raise two research questions (RQ).

RQ1: To what extent does students' understanding of the nature of models and the purpose of models vary across different biological contexts?

RQ2: How frequently are the three levels of understanding of the nature of models and the purpose of models represented among students of different grades?

On the basis of findings which indicate that students' understanding of models in biology is context-specific (Al-Balushi, 2011; Krell et al., 2012, 2014a; Lee et al., 2015; Pluta et al., 2011), we expect an influence of the context on the students' level of understanding. Some contexts might evoke a more elaborate understanding of models and modeling than others (cf. Fischer et al., 1993; Kitchener et al., 1993).

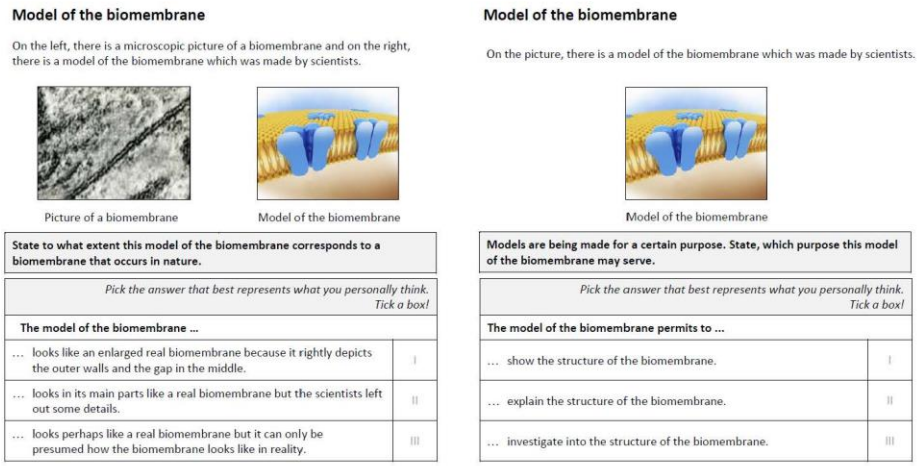
Concerning the differences between grades, research literature permits to formulate two hypotheses: Students' understanding of models and modeling (H1) increases across grades (e.g., Al-Balushi, 2011; Chittleborough et al., 2005), (H2) does not change across grades (e.g., Patzke et al., 2015; Treagust et al., 2002).

Methods

Instrumentation

In order to diagnose students' understanding, we used sets of forced choice tasks for the aspects 'nature of models' and 'purpose of models' respectively. The tasks were developed on the basis of the theoretical framework of model competence and on students' responses to open-ended diagnostic tasks (Grünkorn et al., 2014). In the forced choice tasks, a student had to choose one out of three given answer options and thereby indicate "which of the [options] included in the item is most indicative of his or her behavior" (McCloy, Heggestad, & Reeve, 2005, p. 225). None of the answer options being wrong distinguishes forced choice tasks from multiple choice tasks. By using forced choice tasks, we gain direct feedback on the students' understanding whilst avoiding tied judgments that often occur with likert-type rating tasks (Böckenholt, 2004). In the tasks for the aspect 'purpose of models', a task stem shows a biological model with a short description, while in the tasks for the aspect 'nature of models', the task stem additionally shows the original which the models refers to (Gogolin & Krüger, 2016a). Due to the observation that students distinguish between models in school and in science (e.g., Gobert et al., 2011; Trier et al., 2014) we included the phrase 'the [model] which was made by scientists'. Figure 1 shows forced choice tasks for the aspects 'nature of

models' and 'purposes of models' and illustrates the three levels of understanding (light grey; cf. Table 1).



Model of the biomembrane

On the left, there is a microscopic picture of a biomembrane and on the right, there is a model of the biomembrane which was made by scientists.

Picture of a biomembrane

Model of the biomembrane

State to what extent this model of the biomembrane corresponds to a biomembrane that occurs in nature.

Pick the answer that best represents what you personally think. Tick a box!

The model of the biomembrane ...

... looks like an enlarged real biomembrane because it rightly depicts the outer walls and the gap in the middle.	I
... looks in its main parts like a real biomembrane but the scientists left out some details.	II
... looks perhaps like a real biomembrane but it can only be presumed how the biomembrane looks like in reality.	III

Model of the biomembrane

On the picture, there is a model of the biomembrane which was made by scientists.

Model of the biomembrane

Models are being made for a certain purpose. State, which purpose this model of the biomembrane may serve.

Pick the answer that best represents what you personally think. Tick a box!

The model of the biomembrane permits to ...

... show the structure of the biomembrane.	I
... explain the structure of the biomembrane.	II
... investigate into the structure of the biomembrane.	III

Figure 1
Forced choice tasks (context bio membrane) for the aspects 'nature of models' (left) and 'purpose of models' (right) from the diagnostic instrument. Image credits for original bio membrane: Klaus Hausmann. Rights were obtained. Image credits for model bio membrane: Maurizio De Angelis. Ion channels. CC BY-NC-ND 2.0. www.flickr.com/photos/wellcomeimages/5814248573

Contextualized tasks, including a variety of biological models (*Tyrannosaurus rex* (TR), Neanderthal man (NT), Jura forest (JF), bio membrane (BM), influenza virus (VS), brain (BR), evolution (EV), water circle (WC), lakeshore zone (LZ)) were developed for the diagnostic instrument since the validity of context-independent approaches of assessing students' perspectives on models (i.e. Grosslight et al., 1991; Treagust et al., 2002) have been questioned (Grünkorn et al., 2014; Krell et al., 2012; Sins et al., 2009). The models were chosen on the basis of previous projects with forced choice tasks (Krell, 2013) and open-ended diagnostic tasks (Grünkorn et al., 2014). Following a student-based typology of biological models (Krell, Upmeier zu Belzen, & Krüger, 2014b), we decided to include a range of biological models on a concrete-abstract continuum that allow to be regarded as both a model *of* something and a model *for* something by students (Gogolin & Krüger, 2016a).

For the purpose of providing valid interpretations of the scores, relevant validity evidence was accumulated in a series of steps. A panel of experts (9 biology education researchers with a focus on models from two universities including the second author) judged that the test content adequately represents the content domain (evidence based on the *test content*, AERA, APA & NCME, 2014). Thinking aloud protocols (Gogolin & Krüger, 2016a) short interviews (Gogolin & Krüger, 2016b) and eye tracking (Gogolin, in prep.) were used as methods to ensure that the cognitive processes underlying students' responses to tasks were indeed meaningful with regard to the construct (evidence based on *response processes*, AERA et al., 2014). Tasks for which this evidence could not be provided were removed (Gogolin & Krüger, 2016a, 2016b). Grade-specific analyses of students' response processes with regard to the final tasks lead us to only approve the tasks for the aspect 'nature of models' for grades eleven and twelve (Gogolin & Krüger, 2016a) and the tasks for the aspect 'purpose of models' for the grades ten to twelve (Gogolin, in prep.) In order to test if the students' responses to the tasks confirm the two-dimensionality (nature of models and purpose of models) on which the proposed test score interpretations are based (evidence based on *internal structure*, AERA et al., 2014), we estimated partial credit models (Masters, 1982) and modelled the aspects as one and two latent dimensions. Both AIC and BIC indices indicated the two-dimensional model to be more appropriate to represent the data (Gogolin, in prep.). Comparisons of tasks within each of the aspects during task development studies revealed context-specific differences (Gogolin & Krüger, 2015, 2016b). Evidence based on *relations to other variables* (AERA et al., 2014) was provided by a study in which the measurement with the forced choice tasks for the aspect 'nature of models' converged with data from open-ended diagnostic tasks and diagnostic interviews where students were likewise asked to state to what extent a biological model corresponds to its original (Gogolin & Krüger, 2016a). The sensibility of the diagnosis was verified by using the instrument to evaluate the success of an intervention study (Gogolin & Krüger, in press).

For the present study, the forced choice tasks were assembled in two questionnaires, one of which contained six tasks referring to the aspect 'nature of models' and the other six tasks referring to the aspect 'purpose of models'. In order to gain qualitative insight into possible reasons for context-specific differences in students' answers, and consequently an insight into which contexts might be better suited to evoke students' upper limit of understanding (Fischer et al., 1993), we pursued an *embedded mixed method design* (Creswell & Plano Clark, 2011) by including two open-ended justification tasks at the end of both questionnaires. In each of these two tasks the students were asked to give reason for their decision in one of the forced

choice tasks: "You just answered six questions concerning different biological models. For the model [X], please give reason for why you chose your answer option. Please also explain why you didn't choose the other answer options." The decision to merely include two open-ended justification tasks rather than having the students justify each of their choices in the forced choice tasks was of pragmatic nature as this procedure would have strained students' resources. Altogether, we collected 600 written responses (20 justifications per forced choice task / per grade).

Sample and data collection

285 students in grades ten to twelve from secondary schools in Berlin (Germany) participated in the study (accidental sampling; Table 2).

Table 2
Demographic data of the sample

		Sex		Age	
Grade	$n_{students}$	♂	♀	Min	Max
10	107	37	70	15	17
11	89	35	54	15	18
12	89	44	45	17	21
∑	285	116	169	-	-

After a standardized instruction informed the students about the procedure, all students were handed a questionnaire. Anonymity was maintained by the creation of a password. There was no time limit on the completion of the questionnaire. The students in grade ten only completed the questionnaire for the aspect 'purpose of models', as previous research had shown that the tasks for the aspect 'nature of models' were not suitable for this grade (Gogolin & Krüger, 2016). The students in grades eleven and twelve first completed the questionnaire for the aspect 'nature of models' followed by the questionnaire for the aspect 'purpose of models'. All in all, 178 students gave answers for the aspect 'nature of models' and 285 students for the aspect 'purpose of models'.

Data analysis

Due to the format of the forced choice tasks, it was possible to directly assign the cross of a student to a level for each of the tasks. Differences in students' understanding across contexts (RQ1) were investigated for different subgroups (grades) using Friedman's ANOVA which tests for differences when there are more than two conditions (contexts) and the same participants (students) have been used in all conditions (Field, 2013). We calculated the reliability coefficient Cronbach's *alpha* as a measure of internal consistency (Field, 2013). In order to gain fruitful information on the reasons for students' choices concerning different contexts, we analyzed the qualitative data from the students' responses to the open-ended justification tasks by means of a *thematic qualitative content analysis* (Kuckartz, 2014) using the qualitative data analysis software MAXQDA 11. Similar students' justifications were grouped together in categories (*inductive approach*, Kuckartz, 2014). This approach is comparable to the approach of Grünkorn et al. (2014), who asked for students' understanding of models and modeling in open-ended diagnostic tasks and who developed a category system which describes students' perspectives within the three levels of the theoretical framework (Upmeier zu Belzen & Krüger, 2010) more precisely. Yet, there are two main differences between the approaches. First, we are not aiming at describing the perspectives (level I, II or III) more closely but at discovering reasons for the choice of one of the perspectives. Second, we did not use open-ended diagnostic tasks but justification tasks which consequently have a strong connection to the before answered forced choice tasks. Therefore, the reason given by the student will depend on the perspective which is inherent to the respective forced choice task. The students' responses were coded by three independent raters with the category system including each category with an explanation and a typical statement by a student. The categories defined reasons for why students chose a certain answer options in regard to a specific context helping us understand the context-specificity of students' answers.

One response by a student may have contained different explanations which would then be coded individually. The first author coded all of the material and the two other raters double coded 50 % of the material. Cohen's Kappa (κ) was calculated to measure the agreement between raters (Brennan & Prediger, 1981; Wirtz & Caspar, 2002). The values were in a good range ($.71 \leq \kappa_{Interrater} \leq .74$; Wirtz & Caspar, 2002), indicating that the category system allows analyzing the data objectively.

In order to investigate the second research question regarding differences between students in different grades, we attributed a general level of understanding for each of the aspects to a

student. Instead of using the mean of the six scores, we calculated the general level by conducting a multiple regression analysis (forced entry) with the median as the dependent variable. This procedure enabled us to respect possible differences of students' understanding across tasks by using them as regression coefficients (each of the tasks becoming a predictor). In order to verify if the results of the multiple regression analysis can be trusted, we checked each of the two regression analyses (one for each aspect) for bias in the model due to the violation of assumptions, for multicollinearity, for the fit of the regression model and for the individual contribution of variables. As a result of the multiple regression analysis, every student had one measure (predicted value), expressing his or her level of understanding (level I, II or III) for the aspect 'purpose of models' and, in grades eleven and twelve, another one for the aspect 'nature of models'. The cross-sectional analysis of differences in students' understanding across grades was performed with the Mann-Whitney test for the aspect 'nature of models' and with the Kruskal-Wallis test for the aspect 'purpose of models', as for this aspect, there were more than two comparisons with independent conditions (Field, 2013).

Findings

Differences in students' understanding across contexts (RQ1)

In order to compare students' understanding across individual contexts, it helps to take a look at the means for students' answers to each of the different tasks per aspect. It becomes obvious that the means differ in the aspect 'nature of models' but they are rather similarly dispersed across contexts in the aspect 'purpose of models' (Figure 2). In the aspect 'nature of models', the context of the evolution followed by the context of the Neanderthal man have the highest means while the contexts of the water cycle and the bio membrane have the lowest means (Figure 2). In the aspect 'purpose of models', the mean is the highest for the context of the *T. rex* and the lowest for the context of the bio membrane.

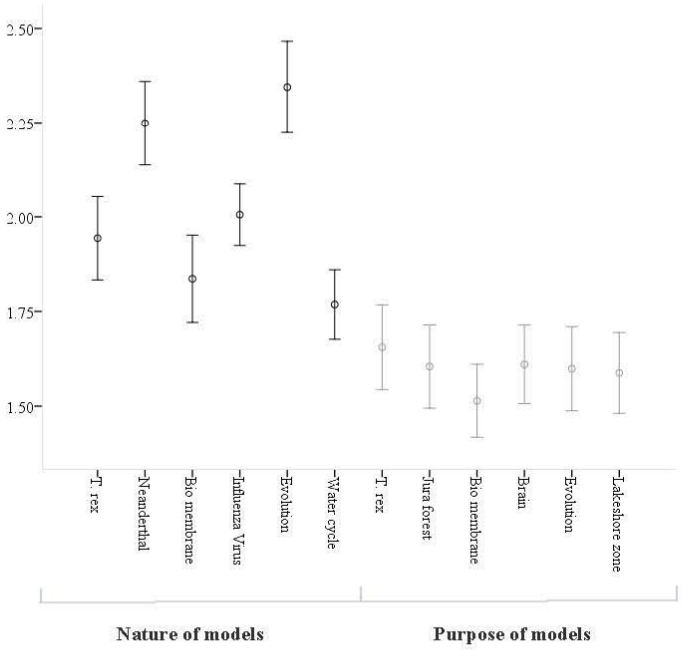


Figure 2
Means and standard errors of the students' preferred levels for each of the models in the aspects 'nature of models' (n=178) and 'purpose of models' (n=285).

We detected low reliabilities between the six tasks for the aspect 'nature of models' ($\alpha_{\text{Grade 11}} = .326$; $\alpha_{\text{Grade 12}} = .311$) in contrast to acceptable reliabilities between the six tasks for the aspect 'purpose of models' ($\alpha_{\text{Grade 10}} = .549$; $\alpha_{\text{Grade 11}} = .696$; $\alpha_{\text{Grade 12}} = .712$). Friedman's ANOVA which tests whether the same student expresses different perspectives across the tasks reveals significant differences across contexts in the aspect 'nature of models' ($\chi^2_{\text{Grade 11}} = 49.5$, $p < .00$, $r = .74$; $\chi^2_{\text{Grade 12}} = 46.9$, $p < .00$, $r = .72$). The same is true for the aspect 'purpose of models' but merely for students in tenth grade ($\chi^2_{\text{Grade 10}} = 68.3$, $p < .00$, $r = .80$; $\chi^2_{\text{Grade 11}} = 5.70$, $p = .34$; $\chi^2_{\text{Grade 12}} = 1.94$, $p = .86$).

The *thematic qualitative content analysis* (Kuckartz, 2014) of students' responses in the open-ended justification tasks brings to light some reasons for the differences in students' choice for a level in the forced choice tasks depending on the context of the task (Tables 3 and 4).

In the aspect 'nature of models', for the context of the bio membrane and the water cycle, students most often reason that the models comply with what is already known about the original (level I) by referring to seemingly trustworthy sources of the knowledge, like teachers or text books (Table 3). For the context of the influenza virus, most students recognized that

in nature, there is a multitude of different viruses which would be impossible to capture in only one model and thus the model comprises average characteristics of all viruses. Consequently, many students chose the answer option on level II indicating that models are idealized representations. The context of the evolution caused the students to comment on the certainty of the knowledge included in the model. Most students state that, due to the complexity and the timely course of the evolution, it is impossible to observe the process directly and to replicate it in a model. The students expressed this uncertainty of the knowledge about the original in general by choosing level III (Table 3).

When justifying their decision in the aspect 'purpose of models', many students gave reason about why a model cannot be used for a certain purpose. We included these explanations into the category system because we find them useful to understand why many students chose answers on level I in the forced choice tasks. In contrast to Figure 2 which shows no considerable differences in the means of the students' preferred levels across the contexts for the aspect 'purpose of models', the content analysis of the students' justifications reveals different reasons across contexts for why students chose a level I answer option. For example, for the context of the Jura forest, the students state that the suggestion to investigate the growth of plants in the Jura forest with the model is not possible because this investigation would need the model to actively change but a model is a snapshot and thus does not change. For the context of the evolution, many students argue that the model cannot be used to predict something because it is an end product of research (Table 4).

Table 3
Students' most frequent justifications for why they chose a certain answer option in the forced choice tasks per context within the aspect 'nature of models'. ($n_{\text{responses per context}}$=40; $N_{\text{total responses}}$=240; n_{codings}=207)

Justification	Example	Context	Chosen answer option "The model [X] …
Model is a replication (level I) because model complies with knowledge.	*I know this model since primary school and that's why I think that it's right.* [sane1108]	BM, WC	… looks like an enlarged biomembrane because it rightly depicts the outer walls and the gap in the middle." … shows the water cycle accurately because scientists used many weather records to build the model."

Model is an idealized representation (level II) because the nature changes over time.	*One cannot be sure whether the influenza virus really looks like that because of constant mutations.* [anna2610]	VS	… accords with a real influenza virus in regard to its shape, but in nature there can be viruses that look different."
Model is an idealized representation (level II) because some parts of the original will never be found.	*Because one can clearly see the long tail from the bones. The rest of the body cannot be determined directly because they have lived millions of years ago.* [siph1003]	NT, TR	… only shows the main features of the Neanderthal man like the bulging eyebrows or the broad nose." … accords in some parts with the real *T. rex*, other parts are not known to the scientists."
Model is a theoretical reconstruction (level III) because there is uncertainty about the knowledge in general.	*Models are products of science and one can present how one thinks that something is until someone else refutes that idea.* [anma2012]	NT, EV	… shows how the human evolution may have taken place, but scientists can only presume." … shows how the Neanderthal man may have looked like, but scientists can only presume."

Note: TR (*Tyrannosaurus rex*), NT (Neanderthal man), BM (bio membrane), VS (influenza virus), EV (evolution), WC (water circle)

Table 4
Students' most frequent justifications for why they chose a certain answer option in the forced choice tasks per context within the aspect 'purpose of models'. ($n_{\text{responses per context}}$=60; $N_{\text{total responses}}$=360; n_{codings}=298)

Justification	Example	Context	Chosen answer option "The model of the [X] serves the purpose to …
Model is used to describe the original (level I) because it can advance one's own knowledge.	*In order to have a picture in my head because I didn't live in that time.* [masa0905]	JF, LZ	… show the basic arrangement of plants in the Jura forest." … show the different lakeshore zones."
Model is used to describe the original (level I) because it makes the original visible.	*The model was created to show the biomembrane which one couldn't see without the model.* [juob2810]	TR, BM, BR, EV	… show the physique of the *T. rex*." … show the structure of the biomembrane." … show the position of the brain areas." … show the descent of the human skull."

Model is not used to explain something about the original ~~(level II)~~ because explanations need to be given as text.	*This model doesn't „explain" anything. In my opinion, explanations are always bound to a text.* [kaie2207]	BM	… explain the structure of the biomembrane."
Model is not used to predict something about the original ~~(level III)~~ because models are end products of research.	*One cannot investigate with the model because a model does only show what has already been discovered.* [yvle0406]	EV	… predict the future development of the human skull."
Model is not used to predict something about the original ~~(level III)~~ because models are too simple for research.	*A further investigation is not possible because it shows only a section* [of the original] *and not all the details are depicted.* [keel1102]	BR, EV	… predict the functions of the brain areas." … predict the future development of the human skull."
Model is not used to predict something about the original ~~(level III)~~ because models are snapshots.	*Development and growth have to be observed over a period of time and for me, models are only snapshots with which one can demonstrate the structure of many things.* [chne1506]	JF	… investigate the growth of plants in the Jura forest."

Note: TR (*Tyrannosaurus rex*), JF (Jura forest), BM (bio membrane), BR (brain), EV (evolution), LZ (lakeshore zone)

Students' understanding across grades (RQ2)

The multiple regression analysis on the basis of the median of the six forced choice tasks resulted in a level for every student for the aspect 'nature of models' and 'purpose of models', respectively. Before presenting the results of the regression analyses we ought to verify the fit of the models to the data.

The examination showed that the assumptions of normality of residuals, linearity and homoscedasticity were met. Multicollinearity could be ruled out because none of the predictors had a strong linear relationship with other predictors ($r < .33$). The Durbin-Watson test showed that the assumption of independent errors is tenable ($1 < d < 3$). The standardized beta values of the predictors certify that each predictor variable made a significant contribution to predicting the outcome ($b > .103$). Regarding the fit of the regression model, the value of R^2 indicated that 62 % of the variability in the outcome of the regression analysis for the aspect 'nature of models' and 77 % for the aspect 'purpose of models' are accounted for by the predictors. Due to the good fit of the multiple regression analysis, we can keep working with the levels attributed to each student for the aspect 'nature of models' and 'purpose of models', respectively.

The majority of students understand models to be idealized representations of their corresponding original (Level II; Figure 3) which serve the purpose to show or to describe the latter (Level I; Figure 3). Nevertheless, the dispersion of levels across grades indicates that with increasing grade, more students are able to recognize that models are theoretical reconstructions which can be used to predict something about an original (Level III). Still, only very few students have these advanced perspectives on models (Figure 3).

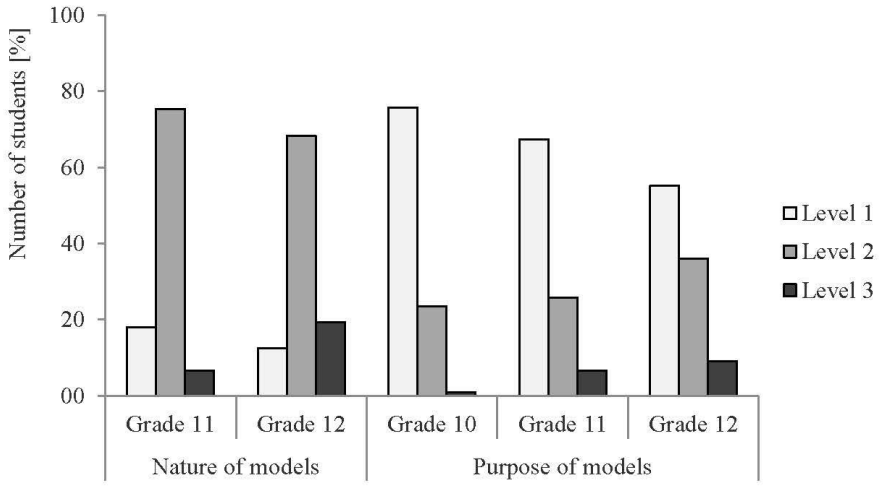


Figure 3
Students' levels of understanding the nature and the purpose of models across grades. *n*=178 for the aspect 'nature of models', *n*=285 for the aspect 'purpose of models'

In the aspect 'nature of models', the levels of understanding of students from grade 11 (*Mdn* = 82.2) were significantly lower than those of students from grade 12 (*Mdn* = 95.9), $U = 3312$, $z = -2.24$, $p = .03$, but with a small effect ($r = 0.2$). The Kruskal-Wallis test also indicated a significant difference across grades for the aspect 'purpose of models' ($H(3) = 11.43$, $p = .01$). We followed up by performing Mann-Whitney tests with a Bonferroni correction which revealed that there were no differences between grades ten and eleven or eleven and twelve but between grades ten and twelve ($U = 3694$, $z = -3.26$, $p = .001$). Again, the effect was small ($r = 0.2$).

Discussion

In this paper, we aim at providing specific and content-rich diagnostic information about students' understanding of the nature and the purpose of models across contexts and across grades as a prerequisite for a successful promotion in the biology classroom (e.g., Campbell et al., 2016; Duit et al., 2012; Günther et al., 2016; Henze et al., 2008). The findings concerning students' understanding of the nature and the purpose of models across contexts and across grades will be judged in relation to findings of other researchers and valuable new insights for teachers will be highlighted. We will also shed light onto the quality of the diagnostic results by discussing methodological limitations of this study.

Students' understanding across contexts (RQ1)

The findings of this study support the claim that students' understanding of models and modeling is context-dependent (Al-Balushi, 2011; Guerra-Ramos, 2012; Krell et al., 2012, 2014a; Leach et al., 2000; Lee et al., 2015; Pluta et al., 2011; Sins et al., 2009). As the comparisons of the means across tasks in the aspect 'nature of models' show, the students more frequently pick answer options on a lower level for the context of the bio membrane and the water cycle and on a higher level in the contexts of the Neanderthal man and the evolution. As for the aspect 'purpose of models' on the other side, the students' understanding overall did not significantly differ across contexts.

Looking at the numbers of the reliability analysis, the low Cronbach's alpha values for the six tasks for the aspect 'nature of models' support the observation that students choose answer options on different levels across the tasks. This result may be interpreted, from a methodical perspective, as a threat to validity as reliability is a requirement for the latter (Field, 2013). On the other side, the numbers may be due to a context-dependency of students' understanding. Leach et al. (2000) state that "Apart from issues of the validity and reliability of the research instruments, a possible interpretation is that students' responses are highly situated and contextualized" (p. 508). Adams and Wieman (2011) pick up on this point by arguing that Cronbach's alpha as an internal consistency coefficient depends on both the correlation between tasks and the number of tasks and thus, having a high correlation between tasks, means that the tasks are repetitive. They go further and argue that in order to create an efficient diagnostic instrument, redundant tasks should be removed. The observation that in the preceding validation analyses the students understood the forced choice tasks as intended

and gave reasonable explanations for their choices (Gogolin, in prep.) makes us believe that the low reliability in the aspect 'nature of models' is a consequence of the students' diverse understanding across the contexts. In order to get more clearly interpretable information about the reliability of the test, it will be necessary to do further investigations. The Standards for Educational and Psychological testing (AERA et al., 2014) suggest three traditional categories of reliability coefficients: (a) alternate-form coefficients, (b) test-retest coefficients and (c) internal-consistency coefficients. Considering that we are suspecting differences in students' understanding across contexts and thus the coefficient values for (a) and (c) could not be clearly interpreted, we will need to use test-retest coefficients which are obtained by administering the same form of the test on separate occasions and then computing Pearson's product-moment correlation coefficient (*r*) or the Intraclass Correlation Coefficient (ICC) (Field, 2013).

Hofer (2006) summarizes the research on context-dependency of epistemic beliefs stating that "few researchers would likely claim that context does not play a role in both shaping and eliciting students' epistemic beliefs, and we have increasing evidence to support this" (p. 90). Krell et al. (2012) point out the importance of context-dependencies in both students' understanding of nature of science as well as the diverse roles which models play in science and conclude that it would be consequential for students' understanding of models and modeling to vary across task contexts. The agreement of science philosophers and science education researchers who see models to be defined only in the context of their use supports this argument (e.g., Bailer-Jones, 2002; Giere, 2001; Mahr, 2008, 2011; Odenbaugh, 2005; van der Valk et al., 2007). Guerra-Ramos (2012) questions the assumption of stable concepts about nature of science in different task contexts and argues that different ideas can be applied in different situations which is why the context matters. This thought is in line with the argumentation by developmental psychologists who see context as part of behavior (Fischer et al., 1993; van Geert, 1998). Students are expected to show better levels of understanding in 'optimal contexts' than in so called 'spontaneous' ones (Fischer et al., 1993). Not only a certain level of understanding but the development of this understanding, according to van Geert (1998), is the effect of interactions between the students' internal properties and the environment. Practicing their understanding in optimal contexts may help students transfer their thinking in these contexts on to other contexts and eventually to develop a more consistent understanding. Clough and Driver (1986) suggest that "once students learn and use a correct scientific explanation in one context they are more likely to employ it in others." (p. 489). The question of what and how these optimal contexts may be cannot be answered

generally. Looking at the students' answers and their justifications for the different contexts used in this study may give us some hints at which contexts might be more 'optimal' than others and consequently which contexts may serve as starting points for promotions concerning models and modeling. In the aspect 'nature of models', the students on average picked the answer options on the highest level in the contexts of the Neanderthal man and the evolution. In the responses from the justification tasks, the students argue that the models of the Neanderthal man and the evolution are theoretical reconstructions because there is uncertainty about the knowledge concerning these phenomena in general. According to the students, this results from both of these models being reconstructions of phenomena that happened partly or completely in a past time which humans have no access to. Such models may be used to introduce the perspective of the nature of models on level III. Still, teachers need to pay attention to contrasting the inability to access knowledge (because time travel has not yet been invented) with the inherent tentativeness of scientific knowledge. It needs to be reflected that the research process has no end and that there is no final state of the knowledge (Lederman & Lederman, 2014). In the aspect 'purpose of models', the means of students' understanding were lower than in the aspect 'nature of models' and students' understanding did not differ significantly across contexts. Nevertheless, the analysis of the qualitative data from the open-ended justification tasks did reveal differences across contexts. These differences referred to the reasons students gave as to why a model cannot serve the purpose to explain or to predict. Some students argue that models are end products of research which show what has been found out already. The perspective may stem from a focus on models as representations which indicate the students' 'science learning performance' (Campbell et al., 2015; Cheng & Lin, 2015). The students critique that the models presented in the tasks are too simple for research as models for research need to include all the details of the original. This corresponds with the perspective of many students that research is only possible with the original but not with the model of the latter. It is likely that students find models useless for the process of research because they ignore the possibility to derive hypotheses from the model which are consequently to be compared with data (Giere, 2001). It underlines the question of whether students generally ignore the importance of hypotheses in scientific research (cf. Carey, Evans, Honda, Jay, & Unger, 1989). After all, we believe that the format of the tasks influenced the students' choice. The qualitative data indicates that students had difficulties imagining the pictures of models in the tasks as concrete physical objects rather than the pictures that they saw. This may inhibit the purposes attributed to the model. For example, many students did not agree with the statement: "The model of the Jura forest serves

the purpose to investigate the growth of plants in the Jura forest." (Jura forest, level III). They argued, that the model of the Jura forest, which was shown to them as a photo taken in the botanical garden, is a snap shot and the growth of plants would need to be observed over a period of time. The diagnosis concerning the purpose of models with our forced choice tasks may not have given optimal contexts to the students in terms of the representational style of the models. In reference to Fischer et al. (1993), we probably offered spontaneous contexts to the students. A diagnosis based on material functional rather than pictorial models may have elicited higher levels within the range of students' understanding.

Concerning assessment, many researchers question the validity of decontextualized approaches to assess students' understanding of models and modeling (e.g., Grünkorn et al., 2014; Krell et al., 2012; Sins et al., 2009) and suggest to develop more contextualized instruments (e.g., Al-Balushi, 2011; Krell et al., 2012; Lee et al., 2015). The results of this study underline the suggestions made by these authors. The results show once again that the influences of the task context need to be carefully considered when interpreting the diagnostic results in order for them to not be a thread to the proposed interpretation of test scores. Fischer et al. (1993) state that "Researchers or educators wanting to assess a child's understanding need to think in terms of a range, not a point on a scale. And they need to always consider context as an integral part of any competence they assess." (p. 104).

Students' understanding across grades (RQ2)

In order to be able to judge upon a student's understanding of the nature and the purpose of models, we assigned one of the three levels of understanding to each student by modeling a multiple regression to our data (Field, 2013). The digitalized version of the diagnostic instrument which was used in this study is currently being used directly by biology teachers with their own students. Thanks to the growing dataset, we will be able to include more predictor variables (e.g., age, school type) in the analysis. Including more variables may improve the fit of the regression model to the data and the accuracy of the prediction (Field, 2013).

The general dispersion of levels shows that the majority of students in all grades, contrary to the pragmatic uses for models in biology outlined by Odenbaugh (2005), see models as idealized representations of an original (level II; nature of models) which have the purpose to show or to describe this original (level I; purpose of models). This observation is in line with

studies describing perspectives of students towards models as rather limited (e.g., Grosslight et al., 1991; Grünkorn et al., 2014).

Students' levels of understanding of the nature and the purpose of models did increase across grades, but with very small effects. These findings are similar to those of other researchers (e.g., Grünkorn et al., 2014; Krell et al., 2014c; Patzke et al., 2015; Treagust et al., 2002). The lack of improvement in students' understanding may result from the way models are implemented in science classrooms. Studies show that models are primarily used as a means to replace or simplify an original in order to transmit information (Danusso et al., 2010; Krell & Krüger, 2016). Furthermore, "distinguishing the positive and negative analogies as clearly as possible" (Hardwicke, 1995, p. 64) as a wish of teachers to ensure that there is no technically false content to be learned, may contribute to the students' understanding of models as idealized representations. Campbell et al.'s (2015) finding that conceptual understanding was the most common pedagogical function identified for modeling adds to this argument. The analyses of Gericke, Hagberg, and Jorde (2013) as well as Ubben, Nitz, Rousseau, and Upmeier zu Belzen (2015) show that the way models are presented in science textbooks is not adapted at establishing modeling as a scientific practice. The argument that science teachers themselves have a rather limited understanding of the nature and the purpose of models (e.g., Borrmann et al., 2014; Crawford & Cullin, 2005; Justi & Gilbert, 2002, 2003; van Driel & Verloop, 1999, 2002) may as well be used to explain the findings. Windschitl and Thompson (2006) argue that "if teachers believe a model is an unproblematic representation of a real-world structure or process, they are less likely to value its development by students or value helping students understand the nature and function of models" (pp. 817–819). Passmore et al. (2014) follow this argument when considering that teachers would be challenged to enact modeling-based curricula.

Rather than focusing on whether or not there is a countable improvement in level of understanding across grades, we can take a grade-specific look at the consistency of students' understanding across contexts. The observation that the Cronbach's alpha values (across the six tasks) increase with grade in the aspect 'purpose of models' may indicate that the students' understanding becomes more consistent. A positive correlation of students' consistent understanding across contexts with the age of the students has been proposed by a number of researchers (e.g., Clough & Driver, 1986; Fischer et al., 1993; King & Kitchener, 2004; Kitchener et al., 1993; Krell et al., 2014c). Clough and Driver (1986) suggest that when students get positive feedback from the teacher concerning an explanation in one context, they are more likely to employ it in others. Ideally, we may add, this explanation is on level III of

the theoretical framework rather than on levels I and II which are being used by students in spontaneous contexts already.

The data of the present study shows that in the highest grade, there are likely to be a number of students who understand models as theoretical reconstructions (nature of models – level III) or who understand that models can be used to derive hypotheses (purpose of models – level III). This more sophisticated understanding of models by some students in higher grades may result from an increase in reflective judgement (King & Kitchener, 2004). King and Kitchener (2004) argue that students in higher grades see knowledge more and more as uncertain and as capturing an element of ambiguity. Lee et al. (2015) see students' grade as one of the potential factors that may impact how students interact with different representations. Students in higher grades are more likely to regard something as a model and consequently score higher in the aspect 'nature of models'. Similar to these results, Al-Balushi (2011) reports that in his study, the overall students' responses to the ECSM survey which was designed to explore students' epistemological positions regarding the credibility of scientific models showed a decrease in the certainty level and an increase in the imaginary level across grades.

Students' more elaborate reflective judgment or their increasing ability to recognize and judge models as representations could build a basis for a successful promotion of students' understanding of models. Knowing that in grades eleven and twelve, there are likely to be some students who already adopted an elaborate understanding of the nature and the purpose of models can be useful for teachers as this knowledge enables them to initiate heterogeneous groups in their classes and thereby integrate the students' preconditions into the class (Duit et al., 2012; Henze et al., 2008). The students who already appreciate the role of models in the process of science can share their perspectives with others. Following Clough and Driver (1986), if teachers were to highlight and praise perspectives of students on level III in some contexts, the students may be able to generalize these perspectives.

The feasibility of these approaches admittedly depends on the diagnostic information the teacher has concerning his or her own class. The instrument which was used in this study could be used by teachers to individually diagnose students while respecting context-specific differences. The forced choice task format allows for an efficient, interpretation-friendly diagnosis and proved sensible enough to evaluate the success of modeling-based interventions (Gogolin & Krüger, 2016a, 2016b, in press).

Conclusion and prospective

Science education standard documents (e.g., Germany: KMK, 2005; UK: QCA, 2007; USA: NGSS Lead States, 2013) declare that the process of thinking in and about models as a form of scientific practice is one of the essential learning goals for science students in order to develop scientific literacy (Hodson, 2014; Oh & Oh, 2011). Consequently, teachers need to be able to foster their students in this domain. Diagnosing their students' understanding of models and modeling constitutes a great challenge for teachers as diagnostic information needs to be both assessed effectively and interpreted validly. Our forced choice tasks provide a possibility for teachers to detect the status quo of their students' understanding of the nature and the purpose of models in biology and consequently foster them more individually also with respect to the different contexts. The results of this study may also serve as a basis to develop interventions to improve teaching and learning about the nature and purpose of models. The reasons given by the students in the open-ended justification tasks can be seen as fruitful entry points for fostering their understanding. For example, in the aspect 'nature of models', often the students seemed ignorant of how the model is built from data. Consequently, teachers may ask their students to construct a model of the water cycle by synthesis modeling (cf. Capps et al., 2016). In this approach, students compare at least two models with distinct surface features but which have a similar underlying structure in order to produce a single, more general model. Hergert et al. (in prep) make students' modeling activity visible with the help of a water black box which produces a specific data pattern. The students are engaged in an active cyclical modeling process where they create and change models based on the data emitted by the black box before finally reflecting about their activities and their models.

The results further indicate that a reflection about models based on different contexts can be helpful because students' understanding of one model may be used to broaden their understanding of other models (Clough & Driver, 1986). A look at the distribution of levels of students' understanding of the nature and the purpose of models across grades emphasizes that a promotion of students understanding is necessary.

References

Adams, W. K., & Wieman, C. E. (2011). Development and Validation of Instruments to Measure Learning of Expert-Like Thinking. *International Journal of Science Education*, *33*(9), 1289–1312.

AERA, APA & NCME [American Educational Research Association, American Psychological Association, & National Council on Measurement in Education] (Eds.). (2014). *Standards for educational and psychological testing*. Washington: American Educational Research Association.

Al-Balushi, S. (2011). Students' evaluation of the credibility of scientific models that represent natural entities and phenomena. *International Journal of Science and Mathematics Education, 9* (3), 571–601.

Bailer-Jones, D. (2002). Scientists' thoughts on scientific models. *Perspectives on Science,10* (3), 275–301.

Böckenholt, U. (2004). Comparative judgments as an alternative to ratings: identifying the scale origin. *Psychological Methods*, *9*(4), 453–465.

Borrmann, J., Reinhardt, N., Krell, M., & Krüger, D. (2014). Perspektiven von Lehrkräften über Modelle in den Naturwissenschaften: Eine generalisierende Replikationsstudie. *Erkenntnisweg Biologiedidaktik*, *13*, 57–72.

Brennan, R., & Prediger, D. (1981). Coefficient Kappa: Some uses, misuses, and alternatives. *Educational and Psychological Measurement*, *41*, 687–699.

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*(1), 159–176.

Campbell, T., Schwarz, C., & Windschitl, M. (2016). What We Call Misconceptions May Be Necessary Stepping-Stones Toward Making Sense of the World. *Science and Children*, *53*(7), 69–74.

Capps, D. K., Shemwell, J., Lindsey, E., Gagnon, L., & Owen, J. (2016). Synthesis Modeling as a Way of Learning through Model Revision. In M. Atwater, M.-H. Chiu, W. Kyle, T. Sondergeld, & R. Yerrick (Eds.), *NARST Annual International Conference Conference Programm Book. Toward Equity and Justice.* (p. 135).

Carey, S., Evans, R., Honda, M., Jay, E., & Unger, C. (1989). 'An experiment is when you try it and see if it works': A study of grade 7 students' understanding of the construction of scientific knowledge. *International Journal of Science Education*, *11*(5), 514–529.

Cartier, J., Rudolph, J., & Stewart, J. (2001). The Nature and Structure of Models in Science. Retrieved from http://biology.westfield.ma.edu/biol104w/sites/default/files/Models.pdf

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*(15), 2453–2475.

Chittleborough, G. D., Treagust, D. D., 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 and Technological Education*, 23, 195–212.

Clough, E. E., & Driver, R. (1986). A study of consistency in the use of students' conceptual frameworks across different task contexts. *Science Education*, *70*(4), 473–496.

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. Eijkelhoff (Eds.), *Research and the quality of science education* (pp. 309–323). Dordrecht: Springer.

Creswell, J., & Plano Clark, V. (2011). *Designing and conducting mixed methods research.* Los Angeles, CA: Sage.

Danusso, L., Testa, I., & Vicentini, M. (2010). Improving Prospective Teachers' Knowledge about Scientific Models and Modelling. *International Journal of Science Education*, *32*(7), 871–905.

Duit, R., Gropengießer, H., Kattmann, U., Komorek, M., & Parchmann, I. (2012). The Model Of Educational Reconstruction – A Framework For Improving Teaching And Learning Science. In D. Jorde & J. Dillon (Eds.), *Science Education Research and Practice in Europe* (pp. 13–37). Rotterdam: Sense Publishers.

Field, A. (2013). *Discovering statistics using IBM SPSS statistics.* Los Angeles: Sage.

Fischer, K. W., Bullock, D., Rotenberg, E. J., & Raya, P. (1993). The dynamics of competence: How context contributes directly to skill. *Development in context: Acting and thinking in specific environments*, *1*, 93–117.

Gericke, N., Hagberg, M., & Jorde, D. (2013). Upper Secondary Students' Understanding of the Use of Multiple Models in Biology Textbooks. *Research in Science Education*, *43*(2), 755–780.

Giere, R. N. (2001). A New Framework for Teaching Scientific Reasoning. *Argumentation*, *15*, 21–33.

Gilbert, J. K., Boulter, C., & Rutherford, M. (1998). Models in explanations, Part 1: Horses for courses? *International Journal of Science Education*, *20*(1), 83–97.

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

Gobert, J., O'Dwyer, L., Horwitz, P., Buckley, B., 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*(5), 653–684.

Gogolin, S. (in prep.) *Diagnosing students' understanding of models and modelling.* Dissertation.

Gogolin, S., & Krüger, D. (2015). Nature of models - Entwicklung von Diagnoseaufgaben. In M. Hammann, J. Mayer, & Wellnitz Nicole (Eds.), *Lehr- und Lernforschung in der Biologiedidaktik* (pp. 27–41). Innsbruck: Studienverlag.

Gogolin, S., & Krüger, D. (2016a). Diagnosing Students' Understanding of the Nature of Models. *Research in Science Education*, 1–23. doi:10.1007/s11165-016-9551-9

Gogolin, S., & Krüger, D. (2016b). Konstruktion von Diagnoseaufgaben zum Zweck von Modellen. *Biologie Lehren und Lernen – Zeitschrift für Didaktik der Biologie*, *20*, 44–62.

Gogolin, S., & Krüger, D. (in press). Modellverstehen im Biologieunterricht diagnostizieren und fördern. *Der mathematische und naturwissenschaftliche Unterricht.* 11 pages.

Grosslight, L., Jay, E., Unger, C., & Smith. (1991). Understanding models and their use in science: Conceptions of middle and high school students and experts. *Journal of Research in Science Teaching*, *28* (9), 799–822.

Grünkorn, J., Upmeier zu Belzen, A., & Krüger, D. (2014). Assessing and structuring students' perspectives on biological models and their use in science to evaluate a theoretical cognitive model. *International Journal of Science Education.* (36), 1651–1684.

Guerra-Ramos, M. T. (2012). Teachers' Ideas About the Nature of Science: A Critical Analysis of Research Approaches and Their Contribution to Pedagogical Practice. *Science & Education*, *21*(5), 631–655.

Günther, S. L., Fleige, J., Upmeier zu Belzen, A., & Krüger, D. (2016). Interventionsstudie mit angehenden Lehrkräften zur Förderung von Modellkompetenz im Unterrichtsfach Biologie. In C. Gräsel & K. Trempler (Eds.), *Entwicklung von Professionalität pädagogischen Personals* (pp. 215–236). Springer Online.

Hardwicke, A. J. (1995). Using Molecular Models to Teach Chemistry. Part 2: Using Models. *School science review*, *77*(279), 47–56.
Henze, I., van Driel, J. H., & Verloop, N. (2008). Development of Experienced Science Teachers' Pedagogical Content Knowledge of Models of the Solar System and the Universe. *International Journal of Science Education*, *30*(10), 1321–1342.
Hergert, S., Krell, M., & Krüger, D. (in prep). How students engage in modelling by using a black box.
Hodson, D. (2014). Learning Science, Learning about Science, Doing Science: Different goals demand different learning methods. *International Journal of Science Education*, *36*(15), 2534–2553.
Hofer, B. K. (2006). Domain specificity of personal epistemology: Resolved questions, persistent issues, new models. *International Journal of Educational Research*, *45*(1-2), 85–95.
Justi, R., & van Driel, J. (2005). The development of science teachers' knowledge on models and modelling. *International Journal of Science Education*, *27*(5), 549–573.
Justi, R., & van Driel, J. (2006). The use of the Interconnected Model of Teacher Professional Growth for understanding the development of science teachers' knowledge on models and modelling. *Teaching and Teacher Education*, *22*(4), 437–450.
Justi, R. S., & Gilbert, J. K. (2002). Science teachers' knowledge about and attitudes towards the use of models and modelling in learning science. *International Journal of Science Education*, *24*(12), 1273–1292.
Justi, R. S., & Gilbert, J. K. (2003). Teachers' views on the nature of models. *International Journal of Science Education*, *25* (11), 1369–1386.
King, P. M., & Kitchener, K. S. (2004). Reflective Judgment: Theory and Research on the Development of Epistemic Assumptions Through Adulthood. *Educational Psychologist*, *39*(1), 5–18.
Kitchener, K. S., Lynch, C. L., Fischer, K. W., & Wood, P. K. (1993). Developmental range of reflective judgment. *Developmental psychology*, *29*(5), 893–906.
KMK [Ständige Konferenz der Kultusminister der Länder in der Bundesrepublik Deutschland]. (2005). Standards für die Lehrerbildung: Bildungswissenschaften.: Beschluss der Kultusministerkonferenz vom 16.12.2004. *Zeitschrift für Pädagogik*, *51*(2), 280–290.
Krell, M. (2013). *Wie Schülerinnen und Schüler biologische Modelle verstehen.* Berlin: Logos.
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*(2), 160–173.
Krell, M., & Krüger, D. (2017). University students' meta-modelling knowledge. Research in Science & Technological Education, 1–13. doi: 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*(3), 367–393.
Krell, M., Upmeier zu Belzen, A., & Krüger, D. (2012b). Students' understanding of the purpose of models in different biological contexts. *International Journal of Biology Education*, *2*(2), 1–34. Retrieved from http://www.ijobed.com/2_2/Moritz-2012.pdf
Krell, M., Upmeier zu Belzen, A., & Krüger, D. (2014a). 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. Part 6. Nicosia, Cyprus: European Science Education Research Association.
Krell, M., Upmeier zu Belzen, A., & Krüger, D. (2014b). How year 7 to year 10 students categorise models. In D. Krüger & M. Ekborg (Eds.), *Research in biological education* (pp. 117–131). Retrieved from http://www.bcp.fu-berlin.de/biologie/arbeitsgruppen/didaktik/eridob_2012/eridob_proceeding/8-How-year.pdf?1389177404
Krell, M., Upmeier zu Belzen, A., & Krüger, D. (2014c). Students' levels of understanding models and modelling in biology: Global or aspect-dependent? *Research in Science Education.* (44), 109–132.
Kuckartz, U. (2014). *Qualitative text analysis*. Washington: Sage.
Leach, J., Millar, R., Ryder, J., & Séré, M.-G. (2000). Epistemological understanding in science learning: the consistency of representations across contexts. *Learning and Instruction*, *10*(6), 497–527.
Lederman, N. G., & Lederman, J. S. (2014). Research on Teaching and Learning of Nature of Science. In N. G. Lederman & S. K. Abell (Eds.), *Handbook of research on science education* (pp. 600–620). New York: Routledge.
Lee, S.-Y., Chang, H.-Y., & Wu, H.-K. (2015). Students' Views of Scientific Models and Modeling: Do Representational Characteristics of Models and Students' Educational Levels Matter? *Research in Science Education*, 1–24. doi:10.1007/s11165-015-9502-x
Mahr, B. (2008). Ein Modell des Modellseins: Ein Beitrag zur Aufklärung des Modellbegriffs. In U. Dirks & E. Knobloch (Eds.), *Modelle* (pp. 187–218). Frankfurt am Main: Peter Lang.
Mahr, B. (2011). On the Epistemology of Models. *Rethinking Epistemology*, *1*, 301–352.
Masters, G. N. (1982). A rasch model for partial credit scoring. *Psychometrika*, *47*(2), 149–174.
McCloy, R. A., Heggestad, E. D., & Reeve, C. L. (2005). A Silk Purse From the Sow's Ear: Retrieving Normative Information From Multidimensional Forced-Choice Items. *Organizational Research Methods*, *8*(2), 222–248.
Nehm, R. H., & Ha, M. (2011). Item feature effects in evolution assessment. *Journal of Research in Science Teaching*, *48*(3), 237–256.
NGSS Lead States. (2013). *Next generation science standards: for states, by states.* Washington: The National Academies Press.
Nicolaou, C., & Constantinou, C. P. (2014). Assessment of the modeling competence: A systematic review and synthesis of empirical research. *Educational Research Review*, *13*, 52–73.
Odenbaugh, J. (2005). Idealized, inaccurate but successful: A pragmatic approach to evaluating models in theoretical ecology. *Biology and Philosophy*, *20* (2), 231–255.
Oh, P. S., & Oh, S. J. (2011). What teachers of science need to know about models: An overview. *International Journal of Science Education*, *33* (8), 1109–1130.
Passmore, C., Gouvea, J., & Giere, R. (2014). Models in science and in learning science. In M. Matthews (Ed.), *International handbook of research in history, philosophy and science teaching* (pp. 1171–1202). Dordrecht: Springer.
Patzke, C., Krüger, D., & Upmeier zu Belzen, A. (2015). Entwicklung von Modellkompetenz im Längsschnitt. In M. Hammann, J. Mayer, & Wellnitz Nicole (Eds.), *Lehr- und Lernforschung in der Biologiedidaktik* (pp. 43–58). Innsbruck: Studienverlag.
Pluta, W. J., Chinn, C. A., & Duncan, R. G. (2011). Learners' epistemic criteria for good scientific models. *Journal of Research in Science Teaching*, *48*(5), 486–511.
QCA [Qualifications and Curriculum Authority]. 2007. *Science: Programme of Study for Key Stage 4.* Retrieved from:

http://media.education.gov.uk/assets/files/pdf/q/science%202007%20programme%20of%20study%20 for%20key%20stage%204.pdf
Schwarz, C. V., Reiser, B. J., Davis, E. A., Kenyon, L., Achér, A., Fortus, D.,. . . Krajcik, J. (2009). Developing a learning progression for scientific modeling: Making scientific modeling accessible and meaningful for learners. *Journal of Research in Science Teaching, 6*, 632–654.
Sins, P., Savelsbergh, E., van Joolingen, W., & 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*(9), 1205–1229.
Treagust, D., Chittleborough, G., & Mamiala, T. (2002). Students' understanding of the role of scientific models in learning science. *Journal of Science Education, 24* (4), 357–368.
Treagust, D., Chittleborough, G., & Mamiala, T. (2004). Students' Understanding of the Descriptive and Predictive Nature of Teaching Models in Organic Chemistry. *Research in Science Education, 34*(1), 1–20.
Trier, U., Krüger, D., & Upmeier zu Belzen, A. (2014). Students' versus Scientists' Conceptions of Models and Modelling. In M. Ekborg, D. Krüger, D. J. Boerwinkel, M. Ergazaki, M. J. Gil Quilez, G. Molinatti et al. (Eds.), *Research in Biological Education* (pp. 103–115). Berlin.
Ubben, I., Nitz, S., Rousseau, M., & Upmeier zu Belzen, A. (2015). Modelle von und für Evolution in Schulbüchern. In U. Gebhard, M. Hammann, & B. Knälmann (Eds.), *Bildung durch Biologieunterricht. 20. Internationale Tagung der Fachsektion Didaktik der Biologie (FDdB) im VBiO* (pp. 75–76). Retrieved from http://www.biodidaktik.de/upload/downloads/1443164473.pdf
Upmeier zu Belzen, A., & Krüger, D. (2010). Modellkompetenz im Biologieunterricht. *Zeitschrift für Didaktik der Naturwissenschaften, 16*, 41–57.
van der Valk, T., van Driel, J. H., & Vos, W. de. (2007). Common Characteristics of Models in Present-day Scientific Practice. *Research in Science Education, 37*(4), 469–488.
van Driel, J. H., & Verloop, N. (1999). Teachers' knowledge of models and modelling in Science. *International Journal of Science Education, 21*(11), 1141–1153.
van Driel, J. H., & Verloop, N. (2002). Experienced teachers' knowledge of teaching and learning of models and modelling in science education. *International Journal of Science Education, 24* (12), 1255–1272.
van Geert, P., & van Dijk, M. (2002). Focus on variability: New tools to study intra-individual variability in developmental data. *Infant Behavior and Development, 25*(4), 340–374.
van Geert, P. (1998). A dynamic systems model of basic developmental mechanisms: Piaget, Vygotsky, and beyond. *Psychological Review, 105*(4), 634–677.
Vo, T., Forbes, C.T., Zangori, L., & Schwarz, C. (2015). Fostering 3rd-grade students' use of scientific models with the water cycle: Elementary teachers' conceptions and practices. *International Journal of Science Education, 37*(15), 2411-2432.
Windschitl, M., & Thompson, J. (2006). Transcending Simple Forms of School Science Investigation. *American Educational Research Journal, 43*(4), 783–835.
Wirtz, M., & Caspar, F. (2002). *Beurteilerübereinstimmung und Beurteilerreliabilität.* Göttingen: Hogrefe.
Zangori, L., Forbes, C.T., & Schwarz, C.V. (2015). Exploring the effect of embedded scaffolding within curricular tasks on 3rd-grade students' model-based explanations about hydrologic cycling. *Science & Education, 24*(7-8), 957-981

Article 5 Mathesius, S. & Gogolin, S. (2017). Die Letzten werden die Ersten sein – Praktisches Modellieren von Planktonkörperformen. *Biologie 5-10, 17,* 10-13.

[Article only accessable as print version or to be purchased online via https://www.friedrich-verlag.de/suche/?q=Reflektierter%20Umgang%20mit%20Modellen]

Article 6 Gogolin, S. & Krüger, D. (in press). Modellverstehen im Biologieunterricht diagnostizieren und fördern. *Der mathematische und naturwissenschaftliche Unterricht*.

[Manuscript as archivable pre-print]

Modellverstehen im Biologieunterricht diagnostizieren und fördern

Sarah Gogolin & Dirk Krüger

Zusammenfassung

Die Förderung von Modellkompetenz als Teil der naturwissenschaftlichen Grundbildung gehört zu den Aufgaben des Biologieunterrichts. Dieser Beitrag stellt zugleich ein Instrument vor, das eine schnelle und aussagekräftige Diagnose von Schülerperspektiven auf Modelle ermöglicht und erläutert Beispiele für Fördermaßnahmen zum Thema Modellkompetenz im Biologieunterricht.

Einleitung

In den Naturwissenschaften sind Modelle wichtige Medien zum Kommunizieren und Beschreiben bereits bekannter Sachverhalte sowie Werkzeuge zum Erforschen unbekannter Phänomene (FLEIGE, SEEGERS, UPMEIER ZU BELZEN & KRÜGER, 2012a). Es geht dabei einerseits darum, dass Modelle Annahmen veranschaulichen, die aktuell plausible Erklärungen für ein Phänomen liefern. Andererseits können aus Modellen auch Vermutungen über ein Phänomen abgeleitet werden. Diese Vermutungen gilt es dann zu überprüfen. So lassen sich mit Modellen ähnlich wie mit anderen wissenschaftlichen Methoden (z. B. Experimenten oder Beobachtungen) neue Erkenntnisse über das untersuchte Phänomen gewinnen. Im Biologieunterricht soll der Einsatz von Modellen in der Wissenschaft reflektiert werden (vgl. KMK, 2005). Diese Arbeit kann durch ein Instrument mit Ankreuzaufgaben unterstützt werden, das eine schnelle und aussagekräftige Diagnose von Schülerperspektiven auf Modelle ermöglicht, indem es die Schüler_innen in drei Niveaus einstuft. Es ermöglicht auf der Basis der Diagnoseergebnisse, Schüler_innen individuell zu fördern. Außerdem kann die Effektivität verschiedener Fördermaßnahmen zur Kompetenzentwicklung beim Denken über Modelle eingeschätzt werden. Das Instrument ist online verfügbar (www.userpage.fu-berlin.de/modelle) und wurde in einer Interventionsstudie im Botanischen Garten Berlin mit Schüler_innen erprobt und evaluiert. Der Beitrag stellt sowohl das Diagnoseinstrument als auch die Stationen der Intervention als Beispiele für Fördermaßnahmen zum Thema Modellkompetenz im Biologieunterricht vor.

Modelle im Biologieunterricht

Die Bildungsstandards für den Mittleren Schulabschluss beschreiben im Kompetenzbereich Erkenntnisgewinnung, dass Schüler_innen Modelle zum Bearbeiten, Veranschaulichen sowie Erklären und Beurteilen komplexer Phänomene nutzen, Modellkritik üben und Modellbildung als wissenschaftliche Denk- und Arbeitsweise anwenden können (KMK, 2005). In der Oberstufe sollen diese Kompetenzen weiter ausgebaut werden. Dabei geht es darum, dass Schüler_innen Modelle nicht nur als Medien zur Unterstützung des Lernens biologischer Fachinhalte verstehen (vgl. HÖGERMANN & CRICKE, 2012a, 2012b), sondern auch als

methodische Werkzeuge, mit denen man die Natur erkunden kann. In dieser methodischen Funktion erlauben Modelle, Hypothesen über die Natur abzuleiten, die eine Prüfung und Untersuchung notwendig machen (vgl. FLEIGE et al., 2012a).

Als Unterstützung bei der Arbeit mit Modellen kann ein Kompetenzmodell für Modellkompetenz dienen, das Perspektiven auf Modelle und die Modellbildung grundsätzlich in fünf Teilkompetenzen beschreibt (KRELL, UPMEIER ZU BELZEN & KRÜGER, in Druck; GRÜNKORN, LOTZ & TERZER, 2014). Hinter der Teilkompetenz „Eigenschaften von Modellen" verbirgt sich die Frage, inwieweit Modelle ihrem biologischen Original entsprechen. Für Schüler_innen, die Modelle häufig als Lernobjekte eines vermeintlich feststehenden Fachinhalts sehen, ist es keineswegs trivial, eine kritische Haltung dem Modell gegenüber zu entwickeln und anzuerkennen, dass Modelle, nicht nur fertige, möglichst genaue Replikationen der Natur sind, deren Aufgabe es ist, Wissensbestände zu transportieren. Man kann Modelle auch als aktuelle Annahmen interpretieren, die einer ständigen Überprüfung standhalten müssen (Tab. 1). In der Teilkompetenz „Alternative Modelle" wird beschrieben, warum es zu einem Original verschiedene Modelle geben kann. Beim „Zweck von Modellen" werden verschiedene Arten des Einsatzes von Modellen dargestellt. Hier drückt sich aus, dass Modelle zum Darstellen und Erklären bekannter Wissensbestände genutzt werden, aber auch zum Erforschen bisher unbekannter Phänomene dienen (Tab. 1). In den Teilkompetenzen „Testen von Modellen" und „Ändern von Modellen" geht es schließlich darum zu überprüfen, ob ein Modell seinen Zweck erfüllt und warum es gegebenenfalls geändert werden muss. Durch eine Unterteilung in Niveaus pro Teilkompetenz bietet das Kompetenzmodell eine Grundlage für eine differenzierte Diagnose und eine entsprechend angepasste Förderung im Biologieunterricht. Das Kompetenzmodell wird hier verkürzt auf die zwei im Diagnoseinstrument enthaltenen Teilkompetenzen dargestellt (Tab. 1).

Tab. 1. Kompetenzmodell für Modellkompetenz (verändert nach KRELL ET AL., in Druck).

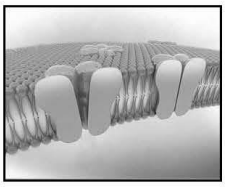
	Niveau I	**Niveau II**	**Niveau III**
Eigenschaften von Modellen	Modelle sind Kopien von etwas	Modelle sind idealisierte Repräsentationen von etwas	Modelle sind theoretische Rekonstruktionen von etwas
Zweck von Modellen	Modellobjekt zur Beschreibung von etwas einsetzen	Bekannte Zusammenhänge und Korrelationen von Variablen im Ausgangsobjekt erklären	Zusammenhänge von Variablen für zukünftige neue Erkenntnisse voraussagen

Modellkompetenz zeigt sich darin, dass die Schüler_innen in allen Teilkompetenzen die drei Perspektiven auf Modelle einnehmen und anwenden können. Studien zufolge werden Modelle im Biologieunterricht vor allem in ihrer Funktion als Medien eingesetzt, um Fachwissen zu vermitteln (BORRMANN, REINHARDT, KRELL & KRÜGER, 2014; CAMPBELL, OH, MAUGHN, KIRIAZIS & ZUWALLACK, 2015). Modelle werden dabei für die Beschreibung und Erklärung von Phänomenen genutzt. In vielen Unterrichtsvorschlägen erfolgt im Rahmen der

Modellkritik häufig ein Abgleich von Modell und Original mit Blick auf die fachliche „Richtigkeit“, um zu vermeiden, dass die Schüler_innen mit dem Modell nicht-angemessene fachliche Inhalte aufbauen (z. B. HÖGERMANN, 2016). Diese überwiegend mediale Nutzung der Modelle bedient die Entwicklung von Perspektiven in Niveaus I und II und erklärt den empirischen Befund, dass Schüler_innen Modelle vornehmlich als Kopien biologischer Strukturen bzw. Prozesse oder als verallgemeinerte Abbilder zur Darstellung und Beschreibung eines Originals einstufen (GROSSLIGHT, UNGER, JAY & SMITH, 1991; GRÜNKORN et al., 2014). Wissenschaftler_innen dagegen nutzen Modelle als Vermutungen, die zwar theoretisch fundiert und plausibel, jedoch nicht sicher sind. Für sie sind Modelle in ihrer voraussagenden Funktion interessant, da ihnen diese Anwendung hilft, neue Erkenntnisse zu erlangen (BAILER-JONES, 2002).

Diagnose von Modellverstehen

Um die Förderung von Modellkompetenz im Biologieunterricht zu unterstützen, wurde ein Diagnoseinstrument entwickelt, das kostenlos online zur Verfügung steht (www.userpage.fu-berlin.de/modelle). In der Online-Version werden die Antworten der Schüler_innen automatisch ausgewertet und es wird anders als in großen Vergleichsstudien sofort eine Rückmeldung über das individuelle Modellverstehen gegeben. Modellverstehen steht für die Fähigkeit, über Modelle und die Modellbildung zu reflektieren (Krell, 2013). Dies stellt neben der Fähigkeit, in problemhaltigen Situationen zu handeln, einen weiteren Bestandteil von Modellkompetenz dar. Beide gilt es im Unterricht zu fördern.
Der Fragebogen besteht aus je sechs geschlossenen Aufgaben für die Teilkompetenzen „Eigenschaften von Modellen“ und „Zweck von Modellen“. In den Aufgaben (z. B. Modell der Biomembran; Abb.1) wählen die Schüler_innen diejenige Perspektive aus, die ihrer Eigenen am ehesten entspricht. Durch verschiedene Studien wurde bestätigt, dass die Aufgaben von Schüler_innen der 11. und 12. Jahrgangsstufe wie intendiert verstanden werden und damit Interpretationen über das Modellverstehen erlauben (GOGOLIN & KRÜGER, 2016; GOGOLIN & KRÜGER, eingereicht).

In der Abbildung siehst du ein Modell der Biomembran, das Biologen entworfen haben.
 Modell der Biomembran
Modelle werden für einen bestimmten Zweck entwickelt. Gib an, <u>welchen Zweck</u> dieses Modell der Biomembran haben kann!

Wähle die Aussage aus, die am ehesten deiner eigenen Meinung entspricht. Mache ein Kreuz!	
Das Modell der Biomembran hat den Zweck ...	
… den Aufbau der Biomembran zu erklären.	II
… den Aufbau der Biomembran weiter zu erforschen.	III
… den Aufbau der Biomembran sichtbar zu machen.	I

Abb. 1. Aufgabenbeispiel zum Modell der Biomembran. Die Niveaus (I, II, III) wurden für den Leser ergänzt.

Die unmittelbare Rückmeldung mit individueller Beschreibung des Modellverstehens kann genutzt werden, um Modellkompetenz differenziert zu fördern. Dies ließe sich realisieren, indem Schüler_innen bezüglich des Modellverstehens in (in)homogenen Gruppen zusammengesetzt werden würden und entsprechende Förderangebote erhielten. Durch die kurze Bearbeitungszeit des Diagnoseinstruments (ca. 10 Minuten) wäre es möglich, das Instrument nach einer Fördermaßnahme ein zweites Mal einzusetzen und damit Veränderungen im Denken über Modelle festzustellen. Die folgenden Vorschläge für Fördermaßnahmen resultieren aus einer Untersuchung, bei der festgestellt werden sollte, ob das Instrument erwartete Veränderungen durch eine Intervention erfassen kann.

Förderung von Modellkompetenz in einer Interventionsstudie

Die Interventionsstudie wurde im Rahmen der Sonderausstellung „modellSCHAU“ im Botanischen Museum Berlin durchgeführt. Im Museum wurden biologische Modelle an Stationen ausgestellt, die mit den Symbolen „Auge“, „Buch“ oder „Glühlampe“ gekennzeichnet waren. Die Symbole standen in Anlehnung an die Teilkompetenz „Zweck von Modellen“ für die Perspektiven „Betrachten“, „Lernen“ und „Forschen“ (GROTZ, 2015).
Das Lernen im Museum kann nach dem *Contextual Model of Learning* (FALK & DIERKING, 2000; vgl. WILDE, 2007) durch eine Reihe von Faktoren aus dem persönlichen, dem soziokulturellen und dem gegenständlichen Kontext charakterisiert werden. Die an fünf aufeinander folgenden Stationen ablaufende Intervention war bis auf eine Station fremdgesteuert. Das Vorwissen der Schüler_innen wurde bei allen Stationen durch kurze Erhebungen der Schülerperspektiven einbezogen (persönlicher Kontext; vgl. WILDE, 2007). Die Erstautorin präsentierte als Vermittlerin außerhalb der Gruppe die Inhalte vor allem in Form kurzer Vorträge und moderierte an einigen Stationen Gespräche zwischen den Schüler_innen (soziokultureller Kontext; vgl. WILDE, 2007). Die Intervention war durch die Führung zu bestimmten Stationen strukturiert und behandelte ausschließlich Stationen mit dem Symbol „Glühlampe“, da diese in besonderem Maße Ansatzpunkte darstellten, um gemeinsam mit den Schüler_innen über den hypothetischen Charakter von Modellen und

deren Einsatz in der wissenschaftlichen Forschung zu reflektieren (gegenständlicher Kontext; vgl. WILDE, 2007).
Insgesamt 65 Schüler_innen aus vier Biologiekursen (Gymnasium, Q-Phase) bearbeiteten vor und nach der 60 minütigen Intervention das Diagnoseinstrument. Als Stationen der Intervention dienten bekannte Modelle aus verschiedenen Themenbereichen der Kursphase im Fach Biologie, die in gleicher Weise auch im Biologieunterricht eingesetzt werden können. Die einzelnen Stationen werden im Folgenden in Bezug auf ihre Eignung zur Förderung von Modellkompetenz und mit weiteren Umsetzungsideen für den Biologieunterricht beschrieben.

Evolution - Rekonstruktion des *Tyrannosaurus rex* (*T. rex*) im Modell

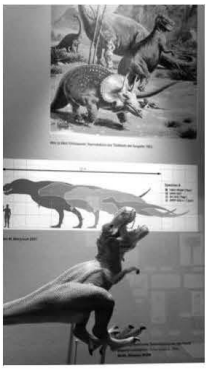


Abb. 2. *T. rex* in aufrechter (oben) und zum Boden paralleler Haltung (unten). Foto aus der Ausstellung.

1865 erstellte Leidy seine Rekonstruktion eines aufrecht stehenden Hadrosaurus. Fünfzig Jahre später enthüllte Osborn im American Museum of Natural History in New York das erste vollständige Skelett eines *T. rex*, in Anlehnung an Leidys Rekonstruktion ebenfalls in aufrechter Haltung. Erst 1970 fanden Wissenschaftler_innen durch Computersimulationen mit einem Modell des *T. rex* heraus, dass eine aufrechte Haltung zu einer Verlagerung der Gelenke geführt hätte. Seitdem wird angenommen, dass der *T. rex* eine zum Boden parallele Haltung einnahm (siehe ROSS, DUGGAN-HAAS & ALLMON, 2013; Abb. 2). Die Frage an die Schüler_innen, welche Eigenschaften eines *T. rex*-Modells auf Spekulation beruhen, führt zu Gesprächen über die Hautfarbe, die Hautoberfläche, die Fortbewegung und die Ernährungsweise des *T. rex*. In diesen Gesprächen wird offensichtlich, dass Modelle keine Kopien einer bekannten Natur sind, sondern auf Annahmen von Wissenschaftler_innen beruhen (Teilkompetenz „Eigenschaften von Modellen"; Niveau III). Konkrete und weitergehende Unterrichtsvorschläge zu Dinosauriern, die es ebenfalls ermöglichen, über den „Wahrheitsgehalt" bzw. die „Richtigkeit" der Rekonstruktionen nachzudenken, finden sich z. B. bei SCHEERSOI und DIERKES (2012) und CHRISTIAN (2012).

Genetik - Die Entwicklung des DNS-Strukturmodells

Ein im Botanischen Museum ausgestelltes DNS-Strukturmodell wurde zum Anlass genommen, um Schüler_innen nach dem Nutzen desselben zu fragen. Häufig waren Schlagworte wie „Veranschaulichung", „Vereinfachung" und „zum Lernen" zu hören. Im Anschluss an diese kleine Meinungserhebung wurden zusammengeschnittene Sequenzen aus dem Kinofilm „Wettlauf zum Ruhm" aus dem Jahr 1987 gezeigt, in dem die jungen Wissenschaftler Watson und Crick ihre Ideen durch ihr Modell immer wieder überprüfen und verändern, ohne je zu wissen, wie die DNS tatsächlich aussieht (siehe BARKE & HARSCH, 2001). Die Schüler_innen wurden erneut nach dem Nutzen des Modells gefragt. Sie nennen nun auch den Zweck, dass Modelle von Wissenschaftler_innen genutzt werden, um etwas Unbekanntes zu erforschen (Teilkompetenz „Zweck von Modellen"; Niveau III). In der Schule können Ausschnitte aus dem Film oder aus dem Buch „Die Doppelhelix" von Watson

(1968) genutzt werden, um mit Schüler_innen ins Gespräch über Modelle als Instrumente der Wissenschaft zu kommen (vgl. KRELL & KRÜGER, 2012; ZABEL, 2001).

Physiologie - Historische Entwicklung des Biomembran-Modells

Ähnlich wie beim DNS-Strukturmodell eignet sich die historische Entwicklung des Biomembran-Modells, um mit den Schüler_innen über die Veränderbarkeit von Wissen und die Rolle von Modellen in diesem Prozess zu reflektieren. Im Museum wurde die Geschichte von DANIELLI und DAVSON und ihrem „Protein-Sandwich-Modell" (1935) nacherzählt (siehe KRELL, HANAUER & FLEIGE, 2012). Besonderer Fokus wurde bei der Erzählung darauf gelegt, dass das Modell von Danielli und Davson von anderen Wissenschaftler_innen stark kritisiert worden war. Nun sollten die Schüler_innen vermuten, wodurch es zur Kritik an einem Modell kommen kann. Die Schüler_innen äußerten hier meist eine Niveau-II-Perspektive zum „Ändern von Modellen", indem sie neue Erkenntnisse als Änderungsgrund anführten. Ein Blick in die Geschichte zeigt dagegen, dass das Modell nicht in der Lage war, den Transport von Stoffen durch die Membran z. B. an Nervenzellen zu erklären und aus diesem Grund geändert werden musste (Teilkompetenz „Ändern von Modellen"; Niveau III). Bis heute werden Transportvorgänge durch Membranen mit Computersimulationen untersucht. Im Unterricht könnten zu diesem Thema z. B. Vorschläge von JAHNKE, AUSTENFELD und LUMER (2013) oder KRAUTWIG (2013) genutzt werden.

Ökologie - *Arabidopsis thaliana* als Modellorganismus

Vor einer Reihe von Töpfchen mit *Arabidopsis thaliana* Pflanzen stehend, wurden die Schüler_innen im Museum gebeten, dieses Ausstellungsstück mit den bisherigen Ausstellungsstücken zu vergleichen. Die Bemerkungen einiger Schüler_innen, es handle sich hierbei nicht wie zuvor um Modelle, sondern um Lebewesen, gaben Anlass, die Begriffe Modell und Original voneinander abzugrenzen. Es wurde herausgearbeitet, dass ein Modell immer zweckgerichtet eingesetzt wird. Die Schüler_innen stellten Vermutungen darüber auf, welchen Zweck die Modellpflanzen für Wissenschaftler_innen haben könnten und nannten Forschungsideen aus den Bereichen „Ökologie" und „Genetik". Statt des im Museum erfolgten, sehr theoretischen Zugangs zu Modellorganismen, können Schüler_innen im Biologieunterricht praktisch mit *Arabidopsis thaliana* arbeiten und eigene kleine Forschungsvorhaben durchführen, wie z. B. die Untersuchung von Dichtestress (siehe GOGOLIN & MATHESIUS, 2014). Dazu werden die Pflanzen in unterschiedlichen Ansätzen angezogen und ihr Wuchs verglichen. Im Rahmen des Forschungsvorhabens können Schritte des naturwissenschaftlichen Erkenntnisprozesses (Fragestellung – Hypothese – Planung und Durchführung – Auswertung; HÖTTECKE & RIEß, 2015; WELLNITZ & MAYER, 2013) reflektiert und kritisch hinterfragt werden. Darüber hinaus gilt es, die Generalisierbarkeit und Übertragbarkeit der Ergebnisse des Experiments mit dem Modellorganismus zu diskutieren. Viele weitere Anregungen zur praktischen Arbeit mit Modellorganismen finden sich bei RUPPERT (2011).

Forschen mit der Blackbox

Abb. 3. Blackbox mit Schnüren.

Den Wissenschaftler_innen nicht nur auf die Hände gucken, sondern selbst Wissenschaftler_in werden! Dieses Potential bietet die Blackbox, deren inneren Aufbau Schüler_innen durch das Ziehen an den Schnüren erforschen sollen (Abb. 3). Im Museum standen die Boxen zur Verfügung. Die Schüler_innen kooperierten im Team, stellten auf der Basis ihrer unterschiedlichen Erfahrungen Modelle vom Inneren der Blackbox auf und suchten nach Möglichkeiten, diese Annahmen durch erneutes Ziehen an den Schnüren zu überprüfen. Letztendlich galt es zu entscheiden, welches Modell das Verhalten der Blackbox zufriedenstellend und widerspruchsfrei erklärt. Hier war – in Analogie zum Erleben in der Wissenschaft – auszuhalten, dass keine Auflösung des Phänomens der Blackbox gegeben wurde. Traumforscher werden auch nie direkt in unsere Köpfe schauen können. Die Arbeit mit der Blackbox motivierte die Schüler_innen, den zirkulären Modellbildungsprozess (Hypothese aus dem Modell ableiten, Testen und Ändern des Modells) mehrfach zu durchlaufen. Zeitgleich zur praktischen Arbeit an der Blackbox reflektierten die Schüler_innen über diesen Prozess. Auch die bleibende Unzufriedenheit in den Wissenschaften, nie ganz sicher sein zu können, ob man wirklich den „richtigen" Mechanismus gefunden hat, wurde von den Schüler_innen akzeptiert. Eine Anleitung für die Arbeit mit der Blackbox im Biologieunterricht wurde z. B. von KRELL und REINISCH (2013) veröffentlicht. Als Onlineressource steht eine „Wasser-Blackbox" in Form eines digitalen Experiments zur Verfügung (HERGERT, 2016; www.tetfolio.fu-berlin.de/web/440484).

Ergebnisse der Intervention und Diskussion zum Einsatz des Diagnoseinstruments

Die Ergebnisse der Intervention im Museum zeigen, dass sich das Modellverstehen der Schüler_innen in beiden Teilkompetenzen in Richtung des Niveaus III entwickelt hat (Abb. 4). Der Unterschied von Pre- zu Posttest ist für beide Teilkompetenzen mit einem großen Effekt statistisch signifikant (Wilcoxon-Test).

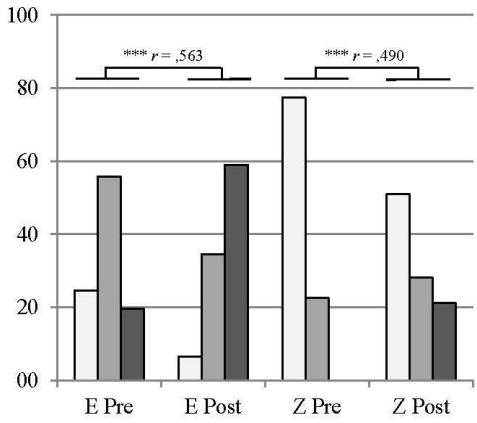


Abb. 4. Ergebnisse des Pre- und Posttests (N=65) in den Teilkompetenzen „Eigenschaften von Modellen“ (E) und „Zweck von Modellen“ (Z). Hellgrau: Niveau I, mittelgrau: Niveau II, dunkelgrau: Niveau III. r = Effektstärke.

Erfreulich sind diese Ergebnisse vor allem vor dem Hintergrund, dass die Intervention nicht als selbstgesteuertes Lernen an einem außerschulischen Standort konzipiert war und damit nicht das Potential dieses besonderen Umfelds ausschöpfte (FALK & DIERKING, 2000). Trotz fehlender vertiefender Erfahrungsmöglichkeiten, fehlender Vorbereitung und Nachbereitung des Besuchs sowie hauptsächlich theoretischer, „lehrerzentrierter“ Elemente, scheint die Intervention erfolgreich gewesen zu sein. Einschränkend muss gesagt werden, dass die Förderung an den Stationen und die Diagnose sowohl zeitlich als auch inhaltlich dicht bei einander lagen und hier nicht von einer dauerhaften Kompetenzförderung ausgegangen werden kann. Trotz aller Bemühungen, Formulierungen aus dem Diagnoseinstrument zu vermeiden, kann nicht ausgeschlossen werden, dass die Diagnoseergebnisse zu einem Teil durch die Sprache der Vermittlerin während der Intervention beeinflusst wurden (*teaching to the test*; MOOSBRUGGER & KELAVA, 2012).

Eine Vorstellungsveränderung zeigt sich besonders deutlich in der Teilkompetenz „Eigenschaften von Modellen“. Modelle werden ähnlich wie in bisherigen Studien (z. B. GROSSLIGHT et al., 1991; GRÜNKORN et al., 2014) zunächst mehrheitlich als Abbilder verstanden, die in den wesentlichen Merkmalen mit ihrem Original übereinstimmen. Mit den Angeboten im Museum wurde zusätzlich die Vorstellung entwickelt, dass Modelle Vermutungen über ein Original darstellen, die auf der Grundlage von theoretischen Erwägungen entwickelt und geprüft sind. Die Intervention bietet an mehreren Stellen (wie z. B. *T. rex*, DNS-Struktur oder Biomembran) Gelegenheiten zu erfahren, dass trotz aller Hinweise aus der Forschung die aktuell wissenschaftlich akzeptierten Interpretationen eines Phänomens immer nur Vorstellungen bleiben, die einen Rest Unsicherheit besitzen (*nature of science*; KOSKA & KRÜGER, 2012). Eine abschließende vollständige Aufklärung ist unter einer naturwissenschaftlichen Perspektive nie zu erreichen. Scheinbar fördert die Einbindung wissenschaftstheoretischer Ansichten und die Auseinandersetzung mit verschiedenen historischen Beispielen in der Teilkompetenz „Eigenschaften von Modellen“ in Richtung des Niveaus III. Viele weitere, potentiell im Unterricht nutzbare historische Beispiele finden sich bei MCCOMAS (2008) oder auf www.storybehindthescience.org.

Für die Teilkompetenz „Zweck von Modellen“ zeigt sich im Pretest, dass die meisten Schüler_innen den Zweck von Modellen in der Darstellung und Beschreibung von bereits bekanntem Wissen sehen (vgl. GROSSLIGHT et al., 1991; GRÜNKORN et al., 2014). Dabei werden Modelle als Medien verstanden, die bekanntes Wissen transportieren. Gründe für die vorherrschend mediale Perspektive auf Modelle liefern Studien zum Einsatz von Modellen im Unterricht (BORRMANN et al., 2014; CAMPBELL et al., 2015). In der Schule dienen Modelle vornehmlich zur Unterstützung beim Lernen von Fachwissen (vgl. HÖGERMANN & CRICKE, 2012a, 2012b) und werden selten als Forschungswerkzeuge beim Erkunden neuer, bisher unbekannter Phänomene, im Unterricht eingesetzt. Auch das für den Biologieunterricht angebotene Lehr- und Lernmaterial in Schulbüchern ist nicht darauf ausgelegt, die Modellbildung als naturwissenschaftliche Erkenntnismethode zu etablieren (UBBEN, NITZ, ROUSSEAU & UPMEIER ZU BELZEN, 2015). Mittlerweile existieren eine Reihe von Themenheften und Beiträgen, die an weiteren Beispielen deutlich machen, wie mit „alt-

bekannten“ Schulmodellen diese Perspektive aufgegriffen werden kann (z. B. FLEIGE et al., 2012a; FLEIGE, SEEGERS, UPMEIER ZU BELZEN & KRÜGER, 2012b; FÜSSENICH et al., 2013). Durch die Intervention konnte in der Teilkompetenz „Zweck von Modellen“ ein Modellverstehen auf Niveau III gefördert werden, jedoch weniger erfolgreich als in der Teilkompetenz „Eigenschaften von Modellen“. Dieser Aspekt muss intensiver, vor allem durch Erfahrungen im Unterricht, gefördert werden. Anregungen aus dem US-amerikanischen Raum fokussieren auf die praktische Arbeit mit Modellen (z. B. LANGE, FORBES, HELM & HARTINGER, 2014; SCHWARZ & WHITE, 2005) und schaffen damit Möglichkeiten, über den Prozess der Modellbildung zu reflektieren. Die Ergebnisse der Interventionsstudie stärken das Vertrauen, dass mit dem Diagnoseinstrument das Modellverstehen in den Teilkompetenzen „Eigenschaften von Modellen“ und „Zweck von Modellen“ gut erfasst werden kann. Dies ermöglicht es, die Vorstellungen der Schüler_innen wiederholt zu diagnostizieren und damit ihre Entwicklung im Biologieunterricht in den beiden Teilkompetenzen zu verfolgen.

Dank

Unser besonderer Dank gilt Christoph van Heteren-Frese für die Umsetzung der Online-Version des Fragebogens.

Literatur

BAILER-JONES, D. (2002). Naturwissenschaftliche Modelle von Epistemologie zu Ontologie. In: A. BECKERMANN & C. NIMTZ (Hrsg.): *Argument und Analyse. Sektionsvorträge, GAP4 e-Proceedings. http://www. gap-im-netz.de/gap4konf/proceedings4/proc.htm* (20.6.2016).

BARKE, H.-D., & HARSCH, G. (2001). Watson und Crick: Nobelpreisträger spielen mit Modellen. In: H.-D. BARKE & G. HARSCH (Hrsg.): *Chemiedidaktik Heute. Lernprozesse in Theorie und Praxis.* Berlin: Springer, 485-496.

BORRMANN, J., REINHARDT, N., KRELL, M., & KRÜGER, D. (2014). Perspektiven von Lehrkräften über Modelle in den Naturwissenschaften: Eine generalisierende Replikationsstudie. *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.

CHRISTIAN, A. (2012). Schlappe Schleicher oder rasante Renner? *Unterricht Biologie, 374*, 42-48.

DANIELLI, J., & DAVSON. H. (1935). A contribution to the theory of permeability of thin films. *Journal of Cellular Comparative Physiology, 5,* 495-508.

FALK, J., & DIERKING, L. (2000). *Learning from Museums.* Walnut Creek: Alta Mira.

FLEIGE, J., SEEGERS, A., UPMEIER ZU BELZEN, A., & KRÜGER, D. (Hrsg.) (2012a). *Modellkompetenz im Biologieunterricht 7-10.* Donauwörth: Auer.

FLEIGE, J., SEEGERS, A., UPMEIER ZU BELZEN, A., & KRÜGER, D. (2012b). Förderung von Modellkompetenz im Biologieunterricht. *Mathematischer und Naturwissenschaftlicher Unterricht*, 65, 19-28.

FÜSSENICH, I. et al. (Hrsg.) (2013). Modellhaft. Aufbau von Modellkompetenz im Sachunterricht. *Grundschule, 6.*

GOGOLIN, S., & KRÜGER, D. (2016). Diagnosing students' understanding of models in biology. *Research in Science Education.* doi: 10.1007/s11165-016-9551-9

GOGOLIN, S., & KRÜGER, D. (eingereicht). *Konstruktion von Diagnoseaufgaben zum Zweck von Modellen.* 18 Seiten.

GOGOLIN, S., & MATHESIUS, S. (2014). Gleich und gleich gesellt sich gern - oder nicht? *Unterricht Biologie, 394,* 21-25.

GROSSLIGHT, L., UNGER, C., JAY, E., & SMITH, C. (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.

GROTZ, K. (Hrsg.) (2015). *Modellschau. Perspektiven auf botanische Modelle.* [Ausstellungskatalog]. Berlin: Laserline.

GRÜNKORN, J., LOTZ, A., & TERZER, E. (2014). Erfassung von Modellkompetenz im Biologieunterricht. *Mathematischer und Naturwissenschaftlicher Unterricht, 67*, 132-138.

HERGERT, S. (2016). *Modelle in der Biologie – Untersuchung einer Blackbox.* Verfügbar unter: www.tetfolio.fu-berlin.de/web/440484.

HÖGERMANN, C. (2016). Primärreaktionen der Fotosynthese. *Mathematischer und Naturwissenschaftlicher Unterricht, 2.2016,* 106-109.

HÖGERMANN, C., & CRICKE, W. (2012a). *Modelle für den Biologieunterricht. Sek. 1.* Wiesbaden: Aulis.

HÖGERMANN, C., & CRICKE, W. (2012b). *Modelle für den Biologieunterricht. Sek. II.* Wiesbaden: Aulis.

HÖTTECKE, D., & RIEß, F. (2015). Naturwissenschaftliches Experimentieren im Lichte der jüngeren Wissenschaftsforschung – Auf der Suche nach einem authentischen Experimentbegriff der Fachdidaktik. *Zeitschrift für Didaktik der Naturwissenschaften, 21*, 127-139.

JACKSON, M. (Produzent & Regisseur). (1987). *Wettlauf zum Ruhm - Die Entschlüsselung der Erbsubstanz* [Film]. Großbritannien: BBC/NDR.

JAHNKE, L., AUSTENFELD, U., & LUMER, J. (2013). Transportvorgänge durch Membranen. *Unterricht Biologie 387/388*, 53-59.

KONFERENZ DER KULTUSMINISTER DER LÄNDER IN DER BUNDESREPUBLIK DEUTSCHLAND (KMK). (2005). *Bildungsstandards im Fach Biologie für den Mittleren Schulabschluss.* München: Wolters Kluwer.

KOSKA, J., & KRÜGER, D. (2012). Nature of Science-Perspektiven von Studierenden. *Erkenntnisweg Biologiedidaktik, 11*, 115–127.

KRAUTWIG, D. (2013). Myelinfiguren. *Unterricht Biologie, 390*, 48-49.

KRELL, M. (2013). *Wie Schülerinnen und Schüler biologische Modelle verstehen.* Berlin: Logos.

KRELL, M., & KRÜGER, D. (2012). Entdeckung der DNS-Struktur. In: J. FLEIGE, A. SEEGERS, A. UPMEIER ZU BELZEN & D. KRÜGER (Hrsg.): *Modellkompetenz im Biologieunterricht 7-10.* Donauwörth: Auer, 49-57.

KRELL, M., & REINISCH, B. (2013). Rätsel um die schwarze Kiste: Mit der Blackbox naturwissenschaftliche Modellbildung verstehen. *Grundschule, 45*, 16-17.

KRELL, M., HANAUER, N., & FLEIGE, J. (2012). Biomembran. In: J. FLEIGE, A. SEEGERS, A. UPMEIER ZU BELZEN & D. KRÜGER (Hrsg.): *Modellkompetenz im Biologieunterricht 7-10.* Donauwörth: Auer, 58-66.

KRELL, M., UPMEIER ZU BELZEN, A., & KRÜGER, D. (in Druck). Modellkompetenz im Biologieunterricht. In: A. SANDMANN & P. SCHMIEMANN (Hrsg.): *Biologie lernen und lehren.* Band 1. Berlin: Logos.

LANGE, K., FORBES, C., HELM, K., & HARTINGER, A. (2014). Forschen heißt auch modellieren! *Sachunterricht, 4*, 17-22.

MCCOMAS, W. (2005). Seeking historical examples to illustrate key aspects of the nature of science. *Science & Education, 17*, 249-263.

MOOSBRUGGER, H., & KELAVA, A. (2012). *Testtheorie und Fragebogenkonstruktion*. Berlin: Springer.

ROSS, R., DUGGAN-HAAS, D., & ALLMON, W. (2013). The Posture of Tyrannosaurus rex. *Journal of Geoscience Education, 61*, 145-160.

RUPPERT, W. (Hrsg.) (2011). Modellorganismen. *Unterricht Biologie, 363*.

SCHEERSOI, A., & DIERKES, P. (2012). Zeig mir deine Zähne, und ich sage dir was du frisst. *Unterricht Biologie, 374*, 12-19.

SCHWARZ, C., & WHITE, B. (2005). Metamodelling knowledge: Developing students' understanding of scientific modelling. *Cognition and Instruction, 23*, 165-205.

UBBEN, I., NITZ, S., ROUSSEAU, M., & UPMEIER ZU BELZEN, A. (2015). Modelle von und für Evolution in Schulbüchern. In: U. GEBHARD, M. HAMMANN, & B. KNÄLMANN (Hrsg.): *Bildung durch Biologieunterricht. 20. Internationale Tagung der Fachsektion Didaktik der Biologie (FDdB) im VBiO* in Hamburg. 75-76.

WATSON, J. (2011). *Die Doppel-Helix: Ein persönlicher Bericht über die Entdeckung der DNS-Struktur*. Berlin: Rowohlt.

WELLNITZ, N., & MAYER, J. (2013). Erkenntnismethoden in der Biologie – Entwicklung und Evaluation eines Kompetenzmodells. *Zeitschrift für Didaktik der Naturwissenschaften, 19,* 315-345.

WILDE, M. (2007). Das Contextual Model of Learning – ein Theorierahmen zur Erfassung von Lernprozessen in Museen. In: D. KRÜGER, & H. VOGT (Hrsg.): *Theorien in der biologiedidaktischen Forschung*. Berlin: Springer, 167-175.

ZABEL, J. (2001): DNA – ein interessantes Spielzeug? Unterrichtsanregung für die Sekundarstufe II. *Unterricht Biologie 268,* 37-43.

Article 7

Gogolin, S., Krell, M., Lange-Schubert, K., Hartinger, A., Upmeier zu Belzen, A., & Krüger, D. (2017). Erfassung von Modellkompetenz bei Grundschüler_innen. In H. Giest, A. Hartinger, & S. Tänzer (Eds.), *Vielperspektivität im Sachunterricht* (pp. 108–115). Bad Heilbrunn: Klinkhardt-Verlag.

[Article only accessable as print version or to be purchased online via http://www.klinkhardt.de/verlagsprogramm/2162.html]