Research in Science Education (2020) 50:2305–2329
https://doi.org/10.1007/s11165-018-9780-1

Assessing Pre-Service Science Teachers' Scientific Reasoning Competencies

Moritz Krell[1] · **Christine Redman**[2] · **Sabrina Mathesius**[1] · **Dirk Krüger**[1] · **Jan van Driel**[2]

Published online: 29 November 2018

Abstract
Scientific reasoning competencies are highlighted in science education policy papers and standard documents in various countries around the world and pre-service science teachers are asked to develop them during teacher training as part of their professional competencies. In order to monitor the development of pre-service science teachers' scientific reasoning competencies during their course of studies and to enable evidence-based improvements of teacher training, instruments for the assessment of scientific reasoning competencies are needed. However, studies propose that the validity of most instruments for assessing scientific reasoning competencies available so far can be questioned. This study presents an English translation of an already developed German multiple-choice instrument to assess pre-service science teachers' scientific reasoning competencies. A sample of $N = 105$ Australian pre-service science teachers participated in this study by answering the translated instrument. Quantitative (differential item functioning, Mantel–Haenszel statistic) and qualitative (think-aloud protocols) analyses provide validity evidence for the translated instrument. Furthermore, the interpretation of the data as an indicator for the participating Australian pre-service science teachers' scientific reasoning competencies suggests that there is a need for a more explicit emphasis on scientific reasoning in Australian science teacher training.

Keywords Scientific reasoning competencies · Pre-service science teachers · Assessment · Test equivalence · Differential item functioning

Introduction

As highlighted in science education policy papers in various countries, science education is vital to "promote a culture of scientific thinking and inspire citizens to use evidence-

✉ Moritz Krell
moritz.krell@fu-berlin.de

[1] Freie Universität Berlin, Berlin, Germany

[2] Melbourne Graduate School of Education, The University of Melbourne, Carlton, Australia

based reasoning for decision making" (European Commission 2015, p. 14). The contribution of science education to develop a society of responsible citizenship was already highlighted in 1966 by the Educational Policies Commission in the United States (p. 16): "What is being advocated here is not the production of more physicists, biologists, or mathematicians, but rather the development of persons whose approach to life as a whole is that of a person who thinks—a rational person" (cf. Lawson 2004). Ever since this has been recognized, the development of abilities related to science and scientific reasoning as well as an understanding of the nature of science is seen as central to enable democratic co-determination in science- and technology-based societies (cf. Driver et al. 1996; Krell et al. 2015a; Lawson 2004) and scientific reasoning competencies (SRC) are seen as summarizing core elements of twenty-first-century skills (Osborne 2013). Consequently, SRC are highlighted in science education policy papers and standard documents in various countries (cf. NGSS Lead States 2013; KMK 2005; VCAA 2016a) and pre-service science teachers are expected to develop them during teacher training as part of their professional competencies (Großschedl et al. 2015; Hartmann et al. 2015; Justi and van Driel 2005; Kleickmann and Anders 2013).

In order to monitor the development of pre-service science teachers' SRC during their course of studies and to enable evidence-based improvements of teacher training, valid instruments for the assessment of these competencies are needed. However, studies propose that the validity of most instruments for assessing SRC available so far can be questioned (Opitz et al. 2017; Osborne 2013). Therefore, and taking the importance of SRC for science education into account, Osborne (2013) calls the development of instruments for assessing SRC "the 21st century challenge for science education."

The present study contributes to this issue by translating a German multiple-choice instrument for assessing pre-service teachers' SRC into English and, by doing so, making it available for use by a broader audience. The original instrument has been evaluated extensively considering various sources of validity evidence (e.g., Hartmann et al. 2015; Mathesius et al. 2018a, b; Stiller et al. 2016), what, taking the criticism on available instruments into account (Opitz et al. 2017; Osborne 2013), can be seen as a distinguishing quality criterion. More precisely, the present study provides evidence for the valid interpretation of the translated instrument's scores as measures of pre-service teachers' SRC and exemplifies its application by measuring a sample of Australian pre-service science teachers' SRC.

Scientific Reasoning Competencies

SRC are understood as a complex construct, which encompass the abilities needed for scientific problem solving as well as to reflect on this process at a meta-level (Lawson 2004; Morris et al. 2012). This includes cognitive processes such as encoding, strategy development, and retrieval (Morris et al. 2012), as well as the application of content knowledge (*knowing that*), procedural knowledge (*knowing how*), and epistemic knowledge (*knowing why*) for problem solving (Kind and Osborne 2017). Content knowledge is understood as knowledge of the domain-specific concepts, procedural knowledge as understanding of scientific procedures and strategies, and epistemic knowledge as knowledge about science and its procedures (Lawson 2004; Morris et al. 2012; Osborne 2013, 2014; White et al. 2011). These knowledge types correspond with three goals of science education highlighted by Hodson (2014): learning science (content knowledge), learning about science (epistemic knowledge), and doing science (procedural knowledge).

Kind and Osborne (2017) further distinguish between six styles of scientific reasoning, each with specific sets of ontological, procedural, and epistemic resources needed for reasoning: mathematical deduction, experimental evaluation, hypothetical modeling, categorization and classification, probabilistic reasoning, and historical-based evaluation. The authors argue that the six styles of reasoning offer a comprehensive schema for the construct of scientific reasoning. Most approaches to operationalize scientific reasoning in science education research focus on one or some of the six styles, for example on modeling (cf. Heijnes et al. 2017; Krell et al. 2017) or experimentation and evidence evaluation (cf. Schauble et al. 1991; van der Graaf et al. 2016). Therefore, Kind and Osborne (2017) identify a lack of clarity and consistency that is related to the construct scientific reasoning in science education research (cf. Opitz et al. 2017). Hence, it is critically important to provide a clear working definition of scientific reasoning and to precisely describe which skills are to be assessed in empirical studies.

The focus of this study is on those two styles of reasoning which are most established in science education research: experimental evaluation and hypothetical modeling (cf. Heijnes et al. 2017; Krell et al. 2017; Schauble et al. 1991; Upmeier zu Belzen and Krüger 2010; van der Graaf et al. 2016; Windschitl et al. 2008). For the operationalization of scientific reasoning competencies in this study, a theoretical framework (Table 1) is used which was developed by combining existing frameworks about scientific reasoning (Mayer 2007) and scientific modeling (Upmeier zu Belzen and Krüger 2010). In this framework, the two sub-competencies *conducting scientific investigations* and *using scientific models* are distinguished, each encompassing different skills (Table 1). The framework describes *conducting scientific investigations* as including scientific experimentation (exploration of causal relationships) and observation (exploration of correlational relationships; Mathesius et al. 2016) and, thus, is broader than *experimental evaluation* (Kind and Osborne 2017), while using scientific models is also understood in a broader sense than *hypothetical modeling* described by Kind and Osborne (2017).

Competencies can be defined as "context-specific cognitive dispositions that are acquired by learning and needed to successfully cope with certain situations or tasks in specific domains" (Klieme et al. 2008, p. 9). Weinert (2001) additionally distinguishes between competencies and skills by proposing to use the term competencies to describe dispositions needed to master tasks with a sufficient degree of complexity, whereas the term skills relates to less complex dispositions. Hence, competencies are "interpreted as a roughly specialized system of individual […] skills that are necessary or sufficient to reach a specific goal" (Weinert 2001, p. 45). This approach of competencies as a system of skills is also applied in several other studies related to the assessment of competencies in educational contexts (e.g., Frey 2006; Krell 2017; Mathesius et al. 2016). However, in this approach, "the boundary between skill and competencies is fuzzy" (Weinert 2001, p. 62).

Table 1 Theoretical framework of this study, including the sub-competencies *conducting scientific investigations* and *using scientific models* (Hartmann et al. 2015; Mathesius et al. 2016)

Scientific reasoning competencies		
Sub-competencies	Conducting scientific investigations	Using scientific models
Skills	Formulating questions Generating hypotheses Planning investigations Analyzing data and drawing conclusions	Judging the purpose of models Testing models Changing models

Following these theoretical approaches (Frey 2006; Klieme et al. 2008; Weinert 2001), it has been assumed that the seven skills *formulating questions*, *generating hypotheses*, *planning investigations*, *analyzing data and drawing conclusions*, *judging the purpose of models*, *testing models*, and *changing models* are necessary to successfully conduct scientific investigations and use scientific models (Hartmann et al. 2015; Mathesius et al. 2016).

Scientific Reasoning Competencies in Science Education and Science Teacher Education

SRC are highlighted in science education policy papers and standard documents in various countries around the world (e.g., NGSS Lead States 2013; KMK 2005; VCAA 2016a). It is assumed that advanced SRC enable democratic co-determination in modern societies and support understanding of science concepts (cf. Driver et al. 1996; Schwarz and White 2005; Thompson et al. 2017). Osborne (2013) further argues that economic needs are one driving force for the integration of SRC in science education standard documents, since they are seen—for example in the PISA studies—as an indicator for the societies' future economic competitiveness (cf. Hanushek and Woessmann 2011; OECD 2010).

As a consequence of the importance of SRC as a learning goal for science education, these competencies are also seen as a goal in science teacher education (Capps and Crawford 2013; Hartmann et al. 2015; Krell and Krüger 2015; Mathesius et al. 2016). In general, teachers' professional competencies encompass (1) professional knowledge; (2) beliefs, values, and goals; (3) motivational orientations; and (4) self-regulation (Baumert and Kunter 2013). Following Shulman (1986), professional knowledge can further be subdivided into pedagogical knowledge (PK), content knowledge (CK), and pedagogical content knowledge (PCK). For science teachers, CK includes epistemic knowledge (Kind and Osborne 2017) related to scientific reasoning; SRC, as the disposition to apply this knowledge for problem solving in scientific contexts, are part of science teachers' professional competencies (Capps and Crawford 2013; Großschedl et al. 2015; Hartmann et al. 2015; Justi and van Driel 2005; Krell and Krüger 2015; Mathesius et al. 2016).

Hence, pre-service science teachers are expected to acquire epistemic knowledge and to develop SRC during their teacher training (Hartmann et al. 2015; Kleickmann and Anders 2013). Evidence from empirical studies suggest that advanced epistemic knowledge may enhance the teaching practice of science teachers (e.g., Capps and Crawford 2013; Krell and Krüger 2015). Consequently, SRC are emphasized in standard documents for science teacher education in many countries around the world (cf. Pedersen et al. 2017).

Assessing Scientific Reasoning Competencies

As a consequence of the importance of SRC for science education, the development of instruments for assessing SRC has become an integral part of science education research (Opitz et al. 2017; Osborne 2013; Stiller et al. 2016). Osborne (2013) identifies the development of instruments for assessing SRC as a core challenge for science education research since assessment instruments may contribute to transform science education for the needs of the twenty-first century. More precisely, Osborne (2013) argues (1) that building assessment instruments helps to clearly operationalize constructs and to communicate the idea of a curriculum, (2) that teachers tend to focus their teaching on constructs that are object of (high stakes) tests, and (3) that assessment items can be used for teaching purposes in science classes. However, Osborne (2013) further states that many assessment instruments related to SRC were developed by psychologists rather than science education researchers and, therefore, follow a domain-general approach. This is not in line with findings about the discipline specificity and knowledge-dependency of scientific reasoning (cf. Krell et al. 2015b). Similarly, Ding et al. (2016, p. 623) criticize that most instruments "are in fact intended to target a broader construct of scientific literacy." In a recent review, Opitz et al. (2017) show that most instruments for assessing scientific reasoning were developed for secondary school students.

Meanwhile, some instruments for assessing scientific reasoning also in higher education have been developed (cf. Hartmann et al. 2015; Mayer et al. 2014). However, the psychometric quality of most published instruments was not evaluated satisfactorily (Opitz et al. 2017) which is why Osborne (2013) criticizes a general lack of validity evidence for the published instruments aimed to assess SRC. This study contributes to fill this gap in science education research by presenting an English translation of an already developed German multiple-choice instrument to assess pre-service science teachers' SRC (cf. Hartmann et al. 2015; Mathesius et al. 2016; Stiller et al. 2016).

Aims of this Study

The aims of this study are to evaluate and to apply a multiple-choice instrument to assess English-speaking pre-service science teachers' SRC. The instrument's application will be exemplified based on a sample of Australian pre-service science teachers.

More precisely, the following research questions (RQ) and associated hypotheses (H) are leading the study:

RQ 1: Evaluation of the instrument.

To what extent does empirical evidence support the validity of the test score interpretation as measures of pre-service science teachers' SRC (cf. AERA et al. 2014; Kane 2013)?

H 1a: It is expected to find psychometric evidence for test equivalence between both versions (i.e., quantitative analysis; Ercikan and Lyons-Thomas 2013) because the multiple-choice instrument is a product of a systematic translation of an existing instrument (based on the TRAPD approach; Harkness 2003; Harkness et al. 2004).

H 1b: It is expected to find evidence based on pre-service teachers' response processes during qualitative pretesting for the validity of the test score interpretation as measures of SRC (i.e., qualitative analysis; Ercikan and Lyons-Thomas 2013) because the multiple-choice instrument is based on the established German instrument (Hartmann et al. 2015; Mathesius et al. 2016).

RQ 2: Application of the instrument.

To what extent does the test score interpretation propose Australian pre-service science teachers possess an adequate level of SRC?

H 2: It is expected that the majority of a sample of Australian pre-service science teachers have a probability higher than 50% for answering the items correctly (i.e., probabilistic analysis; cf. Krell et al. 2015a), because Australian pre-service science teachers are asked to develop SRC during teacher training (e.g., Won et al. 2017).

In the following section, the integration of SRC in science teacher education both in the context of original test development (Germany) and in the context of evaluation and application of the translated English version (Australia) are explained.

Scientific Reasoning Competencies in Science Teacher Education in Germany and Australia

Since the year 2004, competencies related to scientific reasoning (e.g., experimenting, modeling) are summarized in one of four broader competence areas to be achieved by all secondary school students in Germany (KMK 2005). For example, students at the end of school year 10 are aimed to be able to plan simple experiments; to conduct experiments and/or to analyze results; to reason the scope and limitation of experimental designs, procedures, and results; and to evaluate the explanatory power of models (KMK 2005, p. 14).

Consequently, science teachers in Germany need to develop those competencies as well, as part of their professional competencies (Großschedl et al. 2015; Kunter et al. 2013). The Standing Conference of the Ministers of Education and Cultural Affairs of the Länder in the Federal Republic of Germany (KMK 2017) have defined that pre-service science teachers at the end of their studies need to be familiar with the inquiry and working methods of their teaching subjects (e.g., biology, chemistry, physics). For example, knowledge about experimenting and modeling is explicitly demanded for pre-service biology teachers (KMK 2017, p. 22). Neumann et al. (2017, p. 40) further emphasize the following knowledge and abilities needed by pre-service biology teachers in Germany:

> display a substantial and expandable biological knowledge, analytical-critical thinking abilities, as well as methodological competencies; are familiar with the epistemology of the natural sciences and are able to apply this knowledge to biological contexts; are skilled in hypotheses-guided experimentation as well as hypotheses-guided comparison and systemizing.

In Australia, secondary school students are asked to develop science inquiry skills related to science understanding. For example, the Victorian Curriculum and Assessment Authority (VCAA 2016b) demands for students at the end of year 10 to be able to formulate questions or hypotheses that can be investigated scientifically; to independently plan, select, and use appropriate investigation types; to construct and use a range of representations, including graphs, keys, models, and formulas; to record and summarize data, as well as to use knowledge of scientific concepts to evaluate investigation conclusions. In addition to this, the Australian Curriculum includes a set of general capabilities, which are defined as “an integrated and interconnected set of knowledge, skills, behaviors and dispositions that can be developed and applied across the curriculum to help students become successful learners, confident and creative individuals and active and informed citizens” (ACARA 2013).

In line with these demands, Australian science teachers are required to implement scientific inquiry approaches in their lessons, including the formulation of research questions, the design of experiments, and the evaluation of evidence as tasks for students (VCAA 2016a). These approaches also aim to enhance students’ general capabilities, such as critical and creative thinking (ACARA 2013). Consequently, knowledge related to scientific reasoning is an important part of science teachers’ professional knowledge in Australia as well (Won et al. 2017). The Australian Professional Standards for Teachers emphasize “know the content and how to teach it" as one standard of teachers' professional knowledge (AITSL 2011), which means that graduate teachers should be able to "demonstrate knowledge and understanding of the concepts, substance and structure of the content and teaching strategies of the teaching area" (AITSL 2011, p. 10). Based on the final draft of the national professional standards for highly accomplished teachers of science published by the Australian Science Teacher Association (ASTA 2009), science teachers both need conceptual knowledge of the scientific disciplines and knowledge about science and scientific reasoning, for example:

> Highly accomplished teachers of science understand the nature and dimensions of science: that it is a body of knowledge, a way of thinking and communicating, asking and answering questions, and of interpreting events and phenomena from scientific perspectives. […] They understand the kinds of questions that can be tested scientifically and those that cannot; how to plan and conduct scientific investigations, collect evidence and interpret data. (ASTA 2009, p. 5).

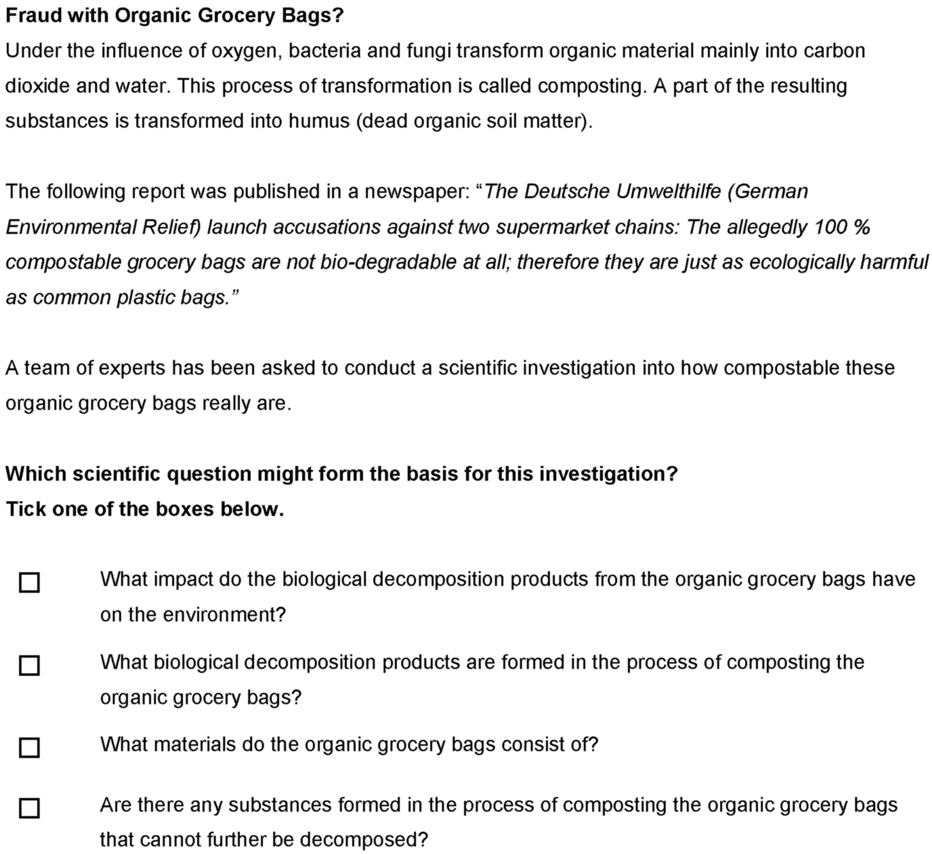
The documents cited above emphasize that science teachers in both Germany and Australia are asked to possess epistemological knowledge about science and its procedures and to have the practical skills and competencies to be able to conduct scientific practices in their science classes. Hence, although not explicitly termed as such, SRC and related diagnostic and teaching competencies are seen as an integral part of science teachers' professional competencies in both Germany and Australia.

Development of the German Instrument in the Ko-WADiS/ValiDiS Project

In order to be able to monitor and to foster the development of pre-service science teachers' SRC throughout their studies as well as to improve science teacher education at the participating universities, the abovementioned multiple-choice instrument was developed in the project Ko-WADiS/ValiDiS (Hartmann et al. 2015; Mathesius et al. 2016; Stiller et al. 2016). The instrument was developed to assess the seven skills described in the theoretical framework shown above (Table 1). Figure 1 exemplarily shows one item related to the skill *formulating questions*.

All items are contextualized in authentic scientific problems and include one attractor (i.e., correct answer) and three distractors (i.e., wrong answers). For the correct identification of the attractor (i.e., for solving the given scientific problem), one needs to apply epistemic knowledge in the respective context. For example, related to the skill *formulating questions* (cf. Fig. 1), one needs to know that scientific questions are related to a phenomenon, are empirically testable, intersubjectively comprehensible, unambiguous, principally answerable, and internally and externally consistent (Mathesius et al. 2014). This operationalization is in line with the definition of competencies as learnable and context-specific dispositions to master complex tasks (Klieme et al. 2008; Weinert 2001) and focuses on the ability to apply epistemic knowledge to solve scientific problems (Mathesius et al. 2014, 2016).

Different sources of validity evidence were considered during the development of the instrument to ensure the plausibility of the test score interpretations as measures of SRC (cf. AERA et al. 2014; Kane 2013). For example, evidence based on test content was considered by using pre-service science teachers' answers to open-ended questions to formulate the attractors and distractors and by conducting test evaluation and revision by experts (Hartmann et al. 2015). Response processes were examined by conducting think-aloud protocols and eye tracking, which confirmed that pre-service science teachers in fact apply epistemic knowledge to solve the multiple-choice items (Mathesius et al. 2018a, b). Evidence based on relation to other



Fraud with Organic Grocery Bags?

Under the influence of oxygen, bacteria and fungi transform organic material mainly into carbon dioxide and water. This process of transformation is called composting. A part of the resulting substances is transformed into humus (dead organic soil matter).

The following report was published in a newspaper: "*The Deutsche Umwelthilfe (German Environmental Relief) launch accusations against two supermarket chains: The allegedly 100 % compostable grocery bags are not bio-degradable at all; therefore they are just as ecologically harmful as common plastic bags.*"

A team of experts has been asked to conduct a scientific investigation into how compostable these organic grocery bags really are.

Which scientific question might form the basis for this investigation?
Tick one of the boxes below.

□ What impact do the biological decomposition products from the organic grocery bags have on the environment?

□ What biological decomposition products are formed in the process of composting the organic grocery bags?

□ What materials do the organic grocery bags consist of?

□ Are there any substances formed in the process of composting the organic grocery bags that cannot further be decomposed?

Fig. 1 One of the three multiple-choice items related to the skill *formulating questions* (item "question 1" in Table 4 and Figure 2). The fourth answering option is the attractor. Note that the figure depicts a translated item in order to allow the readers of the article to understand its content. The full instrument is available for interested readers upon request per email (see Mathesius et al. (2018b) for the German version of this item).

variables was obtained by showing that pre-service biology teachers studying another science subject outperformed those who study a non-scientific subject (e.g., linguistics or social sciences) next to biology (Mathesius et al. 2016) and that pre-service science teachers' test scores increase during the course of studies (Hartmann et al. 2015; Mathesius et al. 2016). Furthermore, statistical analyses revealed good psychometric properties (e.g., item fit parameters, reliability measures) and a one-dimensional data structure representing one overall latent dimension *scientific reasoning* (Hartmann et al. 2015; Mathesius et al. 2016).

Since some authors criticize a lack of validity evidence for instruments to assess SRC (Opitz et al. 2017; Osborne 2013), the consideration of this comprehensive validity evidence can be seen as a distinguishing quality criterion of the instrument. The instrument is currently applied at 11 universities in the German-speaking countries Germany and Austria (currently $N \approx 8500$; cf. Hartmann et al. 2015) to monitor pre-service science teachers' SRC.

Because of missing instruments in other languages and to enable international comparisons, we translated a short version of the instrument (21 multiple-choice items, three items for each skill; Table 1) into English (procedure see below). This short version has already been translated into Spanish and is applied at several universities in Chile, South America (cf. Krell et al. 2018).

Translation of the German Instrument into English

The translation of the instrument followed the *Translation, Review, Adjudication, Pretesting, and Documentation* approach (TRAPD; Harkness 2003; Harkness et al. 2004). The TRAPD approach is established in the social and educational sciences (cf. Forsyth et al. 2016) to warrant valid translations of testing instruments, i.e., to ensure test equivalence between the original and the translated version (Ercikan and Lyons-Thomas 2013; Harkness 2003).

The German instrument was translated into English independently by two translators (*parallel translation*; Harkness et al. 2004). One translator was a professional translator with English being his first language, who already had experience in translating science education research articles from German into English. The other translator learned English as his second language (German as the first) during several stays in English-speaking countries and studied biology and chemistry for becoming a secondary science teacher. Hence, both translators had sufficient expertise in the German and the English language as well as in science education (cf. Harkness et al. 2004). The translators were asked to focus more on the meaning of the information units (i.e., task contexts, questions, attractors, distractors) than on the direct translation of single words (Harkness et al. 2004). However, formal item features should be considered as well (e.g., text length; Stiller et al. 2016) and key terms related to scientific inquiry, such as hypothesis or research question, should be kept and translated directly.

In the review phase, both translations were compared and discussed within a committee consisting of one translator, three authors of this article and three additional science education researchers literate in scientific reasoning. Hence, as a whole, the committee had good skills in the English and the German language, and was familiar with the development of assessment instruments in science education and the underlying construct (cf. Harkness 2003). General comments on both versions have been recorded and differences between the versions have been compared and discussed. Based on these records and discussions, an English version of the instrument was developed by the first author of this article being the *adjudicating body* (Harkness 2003), in cooperation with single members of the reviewing committee; "adjudicators carry the final responsibility for the translation decisions but need not have the translation skills required of reviewer and translators" (Harkness et al. 2004, p. 465). This systematic translation procedure aimed to contribute to test equivalence between the German and the English version of the instrument (Ercikan and Lyons-Thomas 2013).

To obtain empirical evidence which support the validity of the test score interpretation as measures of pre-service science teachers' SRC (RQ 1), the English version of the instrument was quantitatively and qualitatively pretested in the target population (Ercikan et al. 2010; Ercikan et al. 2004; Ercikan and Lyons-Thomas 2013; Harkness 2003; Harkness et al. 2004).

Evaluation of the Instrument (RQ 1): Methods

Quantitative Pretesting: Data Collection, Sample, Data Analysis

Data collection: For data collection, the English version of the instrument was administered in the way it was administered in the German context (equivalence of testing conditions; Ercikan and Lyons-Thomas 2013). The administration took place in regular postgraduate science courses for pre-service science teachers at the graduate school of education of a university in Victoria, Australia. Ethics approval was obtained from the local Human Ethics Approval Group. The standardized test instruction (translated into English) used in the Ko-WADiS/

ValiDiS project was applied (Hartmann et al. 2015), which entails short information about the study and the assessed competencies, and highlights that participation is anonymously and voluntarily.

Sample: In sum, $N = 105$ pre-service science teachers enrolled in a Master of Teaching participated in this part of the study (mean age was 27 years). Table 2 provides further information about the sample.

Data analysis: Data analysis was mainly done within the framework of item response theory (cf. Bond and Fox 2001; Embretson and Reise 2000) using the software ACER Conquest 3 (Wu et al. 2007). The One-Parameter Logistic Model (1PLM) for dichotomous items ("Rasch Model") was applied and the marginal maximum-likelihood estimator was used. In the 1PLM, the probability p of a correct answer for person s on item i is estimated as follows (Embretson and Reise 2000):

$$P(X_{is}) = \exp(\theta_s - \beta_i)/1 + \exp(\theta_s - \beta_i)$$

In this equation, θ_s is the person's trait level and β_i is the item's difficulty. Before interpreting the estimated parameters, the fit between data and model has to be evaluated (Burnham and Anderson 2004). For this purpose, the sum of squared standardized residuals (MNSQ) is proposed. The MNSQ has an expected value of 1 with acceptable values ranging from 0.8 to 1.2 (Bond and Fox 2001). Conquest computes a weighted and an unweighted MNSQ. Because the unweighted MNSQ is more sensitive to outliers than the weighted MNSQ, both statistics should be considered. In addition, t-standardized fit statistics based on the MNSQ are provided, which should range from - 2 to 2 (Wu et al. 2007).

As it was done in the Ko-WADiS/ValiDiS project (Mathesius et al. 2016), both a one-dimensional (*scientific reasoning*) and a two-dimensional (*conducting scientific investigations* and *using scientific models*) 1PLM was specified. The models have been compared based on the information indices AIC and BIC and using the likelihood difference (LD) test (Burnham and Anderson 2004; Wu et al. 2007).

Differential item functioning (DIF) was analyzed to provide psychometric evidence for test equivalence between the English and the German version of the instrument (RQ 1). For this analysis, those German pre-service science teachers who answered the German short version of the instrument were considered ($N = 610$; $n_{\text{Bachelor students}} = 126$; $n_{\text{Master students}} = 109$; $n_{\text{other/did not specify course of study}} = 288$; 80% female). For analyzing DIF, the Mantel–Haenszel (MH) statistic was estimated (Wu et al. 2007; Zwick et al. 1999). This approach is suitable to compare the probability of answering an item correctly between two (or more) matched groups, depending on their test performance. In this study, two matched groups (i.e., German and Australian pre-service science teachers) have been defined based on weighted likelihood estimates (WLE) as ability measures (cf. Wu et al. 2007).

Table 2 Sample information

	Study program			
	Early childhood teacher	Primary school teacher	Secondary school teacher	
Number of students	$n_{\text{female}} = 12$ $n_{\text{male}} = 1$ $n = 13$	$n_{\text{female}} = 44$ $n_{\text{male}} = 11$ $n = 55$	$n_{\text{female}} = 20$ $n_{\text{male}} = 17$ $n = 37$	$n_{\text{female}} = 76$ $n_{\text{male}} = 29$ $N = 105$

Qualitative Pretesting: Data Collection, Sample, Data Analysis

Data collection: For the qualitative pretesting of the English version of the instrument (RQ 1), think-aloud protocols with (primarily) concurrent verbalizations have been conducted in individual interviews (Ercikan et al. 2010; Ericsson and Simon 1998; Harkness et al. 2004; Roth et al. 2013). Hence, pre-service science teachers were asked to work on the instrument and to directly verbalize their thinking (Ericsson and Simon 1998). If necessary, the first author, who was present in all interviews, asked additional clarifying questions (retrospective verbalizations; Ericsson and Simon 1998).

For participating in the interviews, the pre-service science teachers received 25 AUD and had the opportunity to enter a prize draw. Due to the cognitive demands of verbalizing, each participant was expected to work on the instrument not longer than about 1 h. Within this time frame, 9 to 18 items have been finished by the participants. The items for the interviews were chosen to ensure that each item was object of an interview at least twice and, based on the findings of the quantitative analysis (see below), a focus was laid on those items with indication for DIF (cf. Ercikan et al. 2010; Roth et al. 2013).

The participants were informed that the purpose of the interviews was to investigate the understandability of the instrument, not to test their competencies. The interviews were audio-taped and fully transcribed verbatim.

Sample: Five pre-service science teachers took part in the interviews, three pre-service secondary school teachers and two pre-service early childhood teachers. The actual interviews lasted between 55 min and 1 h 14 min. All participants were studying at the graduate school of education mentioned above and ethics approval was obtained for this part of the study as well.

Data analysis: Data analysis was done within the methodological frame of qualitative content analysis (Schreier 2012), based on an already available coding scheme which was developed and applied in the Ko-WADiS/ValiDiS project to analyze think aloud protocols of pre-service science teachers working on the German instrument (Mathesius et al. 2018b). Hence, in this study, the category system was mainly deductively applied to analyze the transcripts (Schreier 2012). However, the original coding scheme was simplified by merging single categories and by including only those categories which have been empirically found in the transcripts.

The resulting category system (Appendix Table 8) entails general categories, for example *reasoned choice of the attractor*, which have been applied independently of the related skill. Due to the focus of this study, one general category was added to grasp verbalizations which indicate language- or culture-related misunderstanding of the items (*lack of text comprehension*/*indication of culture-related misunderstanding*). Furthermore, skill-specific categories describe explanations provided by the respondents, which are specific for the skills.

The coding of the data was done independently by two authors of this article. Cohen's Kappa (κ) was calculated as a measure for intercoder agreement (Brennan and Prediger 1981); disagreements were resolved by discussion after calculating Kappa.

Application of the Instrument (RQ 2): Methods

Data of the sample described above (Table 2) have been analyzed to discuss RQ 2. The software ACER Conquest 3 (Wu et al. 2007) was used to estimate person abilities (θ_s) and item difficulties (β_i) at the latent level based on the 1PLM. As a consequence of the 1PLM's equation (see above), a person with $\theta_s = \beta_i$ has a probability of 50% to answer the respective item correctly; if $\theta_s > \beta_i$, this probability is more than 50%, and if $\theta_s < \beta_i$, this probability is less than 50% (Bond and Fox 2001). WLE were used as estimates for θ_s (cf. Wu et al. 2007), i.e., as indicators for the pre-service science teachers' SRC.

To evaluate to what extent the data propose the pre-service science teachers to possess an adequate level of SRC (RQ 2), the relation of person ability and item difficulty as described above has been used. More precisely, three subgroups of the whole sample have been distinguished for each skill of scientific reasoning (Table 1): respondents with WLE > M_β + 1 SD_β were defined as having advanced competencies, whereas respondents with WLE < M_β - 1 SD_β were labeled as having basic competencies. Respondents who have somewhere between advanced and basic competencies (i.e., WLE = $M_\beta \pm 1$ SD_β) were referred to as transitional.

Evaluation of the Instrument (RQ 1): Findings

Quantitative Pretesting

The dimensionality analysis results in no clear preference for the 1D or the 2D model. Based on the information indices AIC and BIC, the 1D model is preferred, but the LD test results in no significant difference between the 1D and 2D model (Table 3). Due to the principle of parsimony (Burnham and Anderson 2004), and as it was done in the Ko-WADiS/ValiDiS project (Hartmann et al. 2015; Mathesius et al. 2016), the 1D model was used for further data analysis.

In the 1D model, the EAP/PV reliability is rel.$_{\text{EAP/PV}}$ = 0.547 and the item separation reliability is rel.$_{\text{Item}}$ = 0.946. MNSQ and corresponding t values indicate a good fit between data and model (unweighted: -0.89 ≤ MNSQ ≤1.14, $|t| \leq 1.00$; weighted: -0.91 ≤ MNSQ ≤ 1.11, $|t| \leq 1.50$).

Within the 1D model, the analyses showed considerable DIF (category "moderate to large") for two items and minor DIF (category "slight to moderate") for one additional item. It is evident that especially items related to the skill *generating hypotheses* seem to be problematic (Table 4). These items are more difficult in the English version than in the German version of the instrument (i.e., MH statistic < 0).

Qualitative Pretesting (RQ 1)

The independent coding of 33% of the data resulted in a very good intercoder agreement (κ = 0.809; cf. Brennan and Prediger 1981; the categories *other* and *reading/paraphrasing the text* have not been considered in this calculation).

Table 5 shows the number of codings in the categories. It is evident that, within the general categories, many verbalizations have been coded as *selection/exclusion without a reason*. This reflects the fact that the pre-service science teachers often gave a short answer like "This seems like the answer" (interview 3) first, but provided a further explanation of their choice afterwards.

Table 3 Dimensionality analysis

Model	Deviance	AIC	BIC	LD test
1D 1PLM	2558	2602	2661	$\chi^2(2) = 0.114$; $p = 0.945$
2D 1PLM	2558	2606	2670	

Table 4 DIF analysis with the Mantel–Haenszel (MH) statistic

	Items	MH statistic[a]	χ^2 test	DIF category[b]
Conducting scientific investigations	Question 1	− 0.771	$\chi^2(2) = 2.326$; $p = 0.313$	Negligible
	Question 2	− 1.106	$\chi^2(2) = 3.156$; $p = 0.206$	Negligible
	Question 3	0.347	$\chi^2(2) = 0.267$; $p = 0.875$	Negligible
	Hypothesis 1	− 1.971	$\chi^2(2) = 8.112$; $p = 0.003$	Moderate to large
	Hypothesis 2	− 1.551	$\chi^2(2) = 7.222$; $p = 0.027$	Moderate to large
	Hypothesis 3	− 0.846	$\chi^2(2) = 2.328$; $p = 0.312$	Negligible
	Planning 1	0.699	$\chi^2(2) = 3.133$; $p = 0.209$	Negligible
	Planning 2	0.997	$\chi^2(2) = 6.218$; $p = 0.045$	Negligible
	Planning 3	0.468	$\chi^2(2) = 0.765$; $p = 0.682$	Negligible
	Conclusion 1	0.573	$\chi^2(2) = 2.124$; $p = 0.346$	Negligible
	Conclusion 2	0.308	$\chi^2(2) = 0.305$; $p = 0.859$	Negligible
	Conclusion 3	− 0.391	$\chi^2(2) = 0.447$; $p = 0.799$	Negligible
Using scientific models	Purpose 1	0.026	$\chi^2(2) = 0.001$; $p = 0.999$	Negligible
	Purpose 2	0.483	$\chi^2(2) = 0.584$; $p = 0.747$	Negligible
	Purpose 3	− 0.453	$\chi^2(2) = 0.401$; $p = 0.818$	Negligible
	Testing 1	1.343	$\chi^2(2) = 7.073$; $p = 0.029$	Slight to moderate
	Testing 2	0.126	$\chi^2(2) = 0.020$; $p = 0.990$	Negligible
	Testing 3	− 0.006	$\chi^2(2) = 0.009$; $p = 0.995$	Negligible
	Changing 1	0.548	$\chi^2(2) = 1.578$; $p = 0.454$	Negligible
	Changing 2	0.890	$\chi^2(2) = 3.706$; $p = 0.157$	Negligible
	Changing 3	0.212	$\chi^2(2) = 0.104$; $p = 0.949$	Negligible

[a] The MH statistic is defined to be negative when an item is more difficult for the focal group (English version) than for the reference group (German version)

[b] “Moderate to large”: $p \leq 0.05$ and |MH statistic| ≥ 1.5, “slight to moderate”: $p \leq 0.05$ and 1.0 < |MH statistic| < 1.5, “negligible”: $p > 0.05$ and/or |MH statistic| ≤ 1.0 (Zwick et al. 1999). Question, hypothesis, planning, conclusion, purpose, testing, and changing indicate the related skill for each item (see Table 1)

The number of codings of the skill-specific categories indicates that the pre-service science teachers specifically reasoned about the skills *formulating questions*, *generating hypotheses*, *planning investigations*, and *testing models*. The items related to the other three skills seem to elicit rather general verbalizations (Table 5).

The category *lack of text comprehension/indication of culture-related misunderstanding* is especially relevant for the focus of the present study. Within this category, nine codings relate to verbalizations which indicate that the pre-service science teachers did not read the task carefully or were not familiar with single terms. For example, one participant did not understand the meaning of “mechanical action” (included in item “hypothesis 1”), another one did not realize that the expressions “grooves on the fingers” and “wrinkled fingers” relate to the same phenomenon (included in item “conclusion 3”). Furthermore, two codings relate to verbalizations which indicate a lack of understanding without providing reasons. More importantly for the present study, five codings hint to problems as a result of the translation process.

- In the item “changing 2”, the scientific problem relates to a model of stem turtles. In this item, the term “carapace” is used in one paragraph without providing its meaning (back shell). However, in the German version of the item, the meaning is provided directly without using the scientific term “carapace”.
- In items “hypotheses 3” and “conclusion 2”, pronouns are used which may have caused ambiguousness of the sentences. For example, in item “hypotheses 3”, the following

Table 5 Number of codings in the categories (description of categories see Appendix Table 8)

	Number of codings
General categories	
Other	90
Reading/paraphrasing the task	157
Reasoned choice of the attractor	22
Lack of text comprehension/indication of culture-related misunderstanding	16
Selection/exclusion based on signal word	10
Selection/exclusion without a reason	57
Categories for the skills *formulating questions* and *generating hypotheses*	
(Includes alternative scientific approach, irrelevant factor, experimental testability impossible, methodological understanding—scientific question/hypothesis, methodological understanding missing—scientific question/hypothesis)	35
Categories for the skill *planning investigations*	
(Includes irrelevant design, methodological understanding—planning investigation, methodological understanding missing—planning investigation)	17
Categories for the skill *analyzing data and drawing conclusions*	
(Includes criticism of method, inappropriate interpretation, methodological understanding—analyzing data/drawing conclusion)	10
Categories for the skill *judging the purpose of models*	
(Includes relevance of described purpose, scientific application of models is denied, criticism of described purpose)	8
Categories for the skill *testing models*	
(Includes alternative scientific approach, testing the model object, no empirical test, criticism of model)	25
Categories for the skill *changing models*	
(Includes alternative scientific approach, model empirically falsified, model not empirically falsified, inappropriate empirical basis)	7

sentence is part of the item stem: “Scientists have hypothesized that the jasmonic acid produced by the plants has a negative signal effect to keep them from further damaging the plant.” One participant was not sure about the meaning of *them* in this sentence (“Well, it’s pretty confusing, keep them from damaging. Keep who? Caterpillars?”; Interview 1). This sentence is clearer written in the German version since caterpillars are directly mentioned and no pronoun is used.

- Finally, two codings relate to the formulation “might form the basis for this investigation”, which is part of the four items “question 1”, “question 2”, “question 3”, and “hypothesis 2”, and is meant in the sense of *underlie this investigations*. For example, in item “question 1”, the participants are asked “Which scientific question might form the basis for this investigation?” and thereafter have to identify the research question which is examined in the investigation described in the item stem (Fig. 1). However, one participant understood the formulation “might form the basis” as something which have to be done *prior* to the investigation (“I think I should be cautious about the meaning of ‘might form the basis.’ So that means if we want to conduct the investigation successfully we have to answer some questions as the basis.”; Interview 4).

Application of the Instrument (RQ 2): Findings

Table 6 indicates the proportion of correct answers for the items of the seven skills and overall based on the raw data (i.e., 1.0 = 100% correct answers). This

Table 6 Proportion of correct answers (*M*) and standard deviation (*SD*) for each skill and overall

		Number of items	*N*	Proportion of correct answers
Conducting scientific investigations	Formulating questions	3	103	$M = 0.29$ ($SD = 0.24$)
	Generating hypotheses	3	103	$M = 0.26$ ($SD = 0.25$)
	Planning investigations	3	103	$M = 0.68$ ($SD = 0.27$)
	Analyzing data and drawing conclusions	3	103	$M = 0.55$ ($SD = 0.25$)
Using scientific models	Judging the purpose of models	3	105	$M = 0.37$ ($SD = 0.28$)
	Testing models	3	104	$M = 0.44$ ($SD = 0.28$)
	Changing models	3	103	$M = 0.52$ ($SD = 0.21$)
Scientific reasoning (i.e., overall)		21	105	$M = 0.45$ ($SD = 0.15$)

data illustrates that the items related to *planning investigations* turned out be rather easy (about 68% correct answers), whereas the items related to *generating hypotheses* were the most difficult ones (about 26%). Overall, the items have been answered correctly by about 45% of the pre-service science teachers.

Figure 2 shows the person ability and item difficulty parameters estimated in the 1PLM within one common logit-scale ("Wright Map"; cf. Bond and Fox 2001). It is evident that the skills *formulating questions* and *generating hypotheses* are rather difficult (i.e., high item difficulty parameters), whereas *planning investigations* is the easiest one (i.e., low item difficulty parameters). The standard deviations illustrate that the difficulty of the items related to *planning investigations* and *analyzing data and drawing conclusions* varies more than those related to the other skills.

The pre-service secondary school teachers performed significantly better in the test ($M_{\mathrm{WLE}} = 0.45$, $SD_{\mathrm{WLE}} = 0.67$) than the pre-service primary school teachers ($M_{\mathrm{WLE}} = -0.22$, $SD_{\mathrm{WLE}} = 0.67$, $p < 0.001$, $d = 0.99$; large effect) and the pre-service early childhood teachers ($M_{\mathrm{WLE}} = -0.23$, $SD_{\mathrm{WLE}} = 0.66$, $p = 0.003$, $d = 1.00$; large effect). There is no significant difference in the WLE between the pre-service primary school teachers and the pre-service early childhood teachers ($p = 0.968$).

Based on the WLE and the definition for basic, transitional, and advanced competencies described above, about 84% ($n = 88$) of the pre-service science teachers have only basic competencies related to *generating hypotheses* (Table 7). In the easiest skill *planning investigations*, all pre-service science

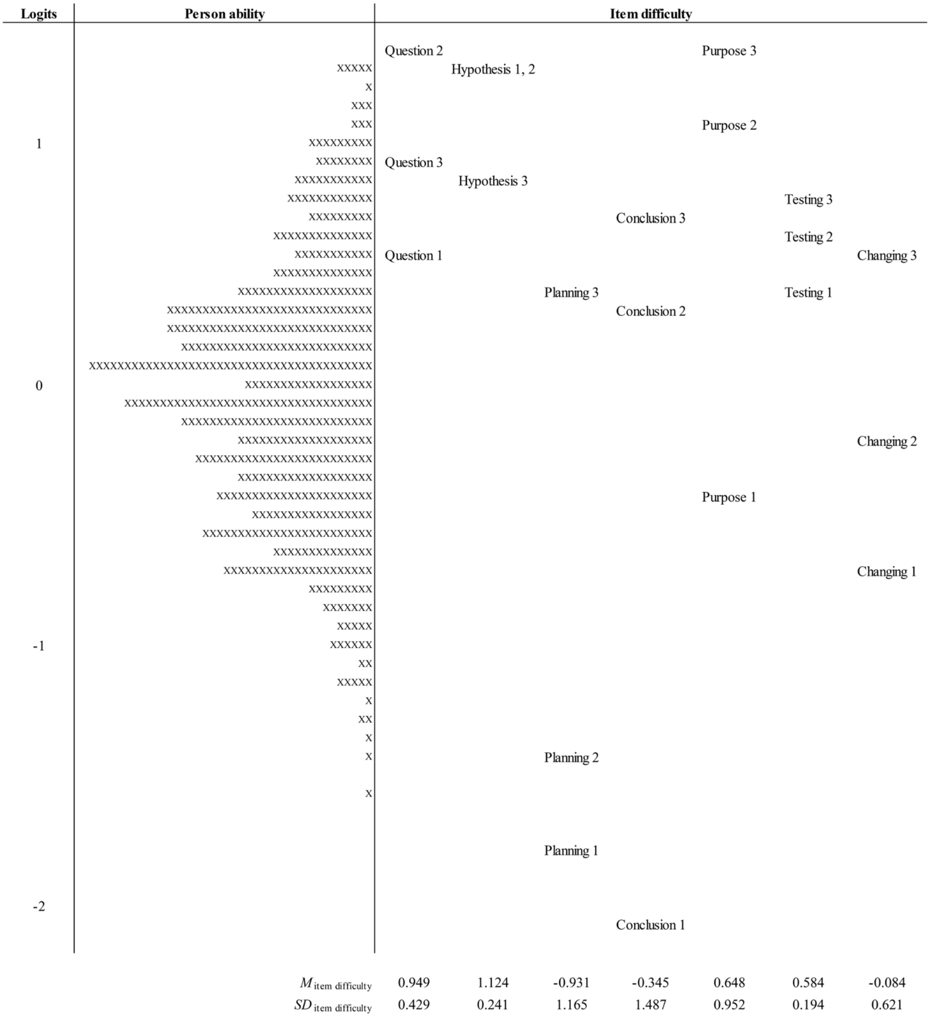


Fig. 2 Wright Map with the person ability and item difficulty parameters estimated in the 1PLM within one common logit-scale. Each X represents 0.2 cases. Question, hypothesis, planning, conclusion, purpose, testing, and changing indicate the related skill for each item (see Table 1)

teachers have either transitional ($n = 63$) or advanced ($n = 42$) competencies. On average, about 40% of the sample have basic, about 47% transitional, and about 13% advanced competencies.

It is evident that pre-service early childhood teachers reached advanced competencies in only two skills (*planning investigations*, *changing models*), whereas pre-service secondary school teachers reached this level in all skills (Table 7).

Table 7 Number of pre-service science teachers within the three subgroups in each skill

			Basic (WLE < M_β − 1 SD_β)	Transitional (WLE = M_β ± 1 SD_β)	Advanced (WLE > M_β + 1 SD_β)
Conducting scientific investigations	Formulating questions	Total sample	81	22	2
		Sec/Pri/Earl	21/49/11	14/6/2	2/–/–
	Generating hypotheses	Total sample	88	15	2
		Sec/Pri/Earl	24/51/13	11/4/–	2/–/–
	Planning investigations	Total sample	–	63	42
		Sec/Pri/Earl	–/–/–	13/39/11	24/16/2
	Analyzing data and drawing conclusions	Total sample	–	97	8
		Sec/Pri/Earl	–/–/–	32/52/13	5/3/–
Using scientific models	Judging the purpose of models	Total sample	33	70	2
		Sec/Pri/Earl	5/24/4	30/31/9	2/–/–
	Testing models	Total sample	72	16	17
		Sec/Pri/Earl	15/46/11	9/5/2	13/4/–
	Changing models	Total sample	20	61	24
		Sec/Pri/Earl	1/16/3	20/33/8	16/6/2

N = 105 (total sample), *Sec* pre-service secondary school teachers (n = 37), *Pri* pre-service primary school teachers (n = 55), *Earl* pre-service early childhood teachers (n = 13)

Discussion and Conclusion

Assessment instruments for scientific reasoning may contribute to transforming science education for the needs for the twenty-first century, for example by clearly operationalizing and communicating the construct and by providing items that can be used for teaching purposes in science classes (Osborne 2013). Furthermore, such instruments enable science education researchers to monitor and to foster the development of pre-service science teachers' SRC throughout their studies and to improve science teacher education (Hartmann et al. 2015; Mathesius et al. 2016).

This study contributes to this issue by presenting an English translation of an established German multiple-choice instrument to assess pre-service science teachers' SRC (Hartmann et al. 2015; Mathesius et al. 2016). The multiple-choice items are contextualized in authentic scientific problems and, thus, take account for the discipline specificity and knowledge dependency of scientific reasoning (Krell et al. 2015b; Osborne 2013). Furthermore, the items relate to the two styles of reasoning, or sub-competencies, respectively, which are most established in science education research: experimental evaluation and hypothetical modeling (cf. Schauble et al. 1991; Upmeier zu Belzen and Krüger 2010; Windschitl et al. 2008). As the underlying construct generally defines and constrains the scope of interpretation of competence assessments (Shavelson 2013), the theoretical framework (Table 1) may be critically discussed as one limiting factor of this study. For example, the focus on two styles of reasoning may be criticized as narrow and reflecting an "impoverished account of scientific thinking" (Kind and Osborne 2017, p. 17) since there are more styles of reasoning used in the sciences (e.g., six styles proposed by Kind and Osborne 2017). Furthermore, the definition and operationalization of the two styles may be critically discussed as well. The approach to define *conducting scientific investigations* distinguishes the important process skills formulating questions, generating hypotheses, planning investigations, analyzing data, and drawing conclusions (Table 1), but omits other processes like actually doing the investigation (e.g., data collection). This limitation, at least partly, is a result of the multiple-choice format and has to be taken into account when interpreting the scores as indicators for pre-service science teachers' SRC (Shavelson 2013). Especially the process of doing an investigation is challenging to assess within a multiple-choice format since most aspects which are important while doing an investigation should be theoretically considered in the planning process as well (e.g., no variable confounding, systematic measurement). Hence, paper-pencil questions about how to do an investigation may rather address skills related to the planning and, therefore, typically omit this skill (e.g., Krell 2017).

As the German version of the instrument was thoroughly evaluated, taking into account various sources for validity evidence (cf. Hartmann et al. 2015; Mathesius et al. 2018b), and the translation of the English version was systematically done based on the TRAPD approach (Harkness 2003; Harkness et al. 2004), this study aimed to obtain psychometric evidence for test equivalence between both versions (H 1a) and evidence based on response processes for the validity of the test score interpretation as measures of SRC (H 1b).

The findings related to H 1a propose the instrument to be widely equivalent to the original version. In line with the analyses of the German data (cf. Mathesius et al. 2016), the English instrument seems to assess a one-dimensional construct since both information indices propose the 1D model to better fit the data than the 2D model (Table 3). The one-dimensional construct assessed by the instrument is likely to be interpreted as *scientific reasoning competencies*. Similar dimensionality of data between test versions are one indicator for test equivalence (Ercikan and Lyons-Thomas 2013). A further indicator, which was evaluated in this study, is analysis of DIF (Ercikan and Lyons-Thomas 2013; Wu et al. 2007; Zwick et al. 1999). The estimated MH statistic proposes only two items, both related to *generating hypotheses*, having considerable DIF (i.e., category "moderate to large"; Table 4). Items within this category "are subjected to further scrutiny and may be eliminated from tests" (Zwick et al. 1999, p. 3). Following this advice, the qualitative analysis of the instrument (H 1b) focused on the items related to the skill *generating hypotheses* (cf. Ercikan et al. 2010; Roth et al. 2013).

The findings of the qualitative analysis of the items ("think-aloud protocols"; Ericsson and Simon 1998) mainly support the assumption that the items elicit epistemic knowledge. This was also found in qualitative analyses of the German version of the instrument (Mathesius et al. 2018b). In the present study, the general category *reasoned choice of the attractor* was identified 22 times among five pre-service teachers, indicating that the pre-service teachers understood the attractor and were able to paraphrase it or to further explain the scientific approach described in the attractor. The skills *formulating questions*, *generating hypotheses*, *planning investigations*, and *testing models* elicited more often skill-specific argumentations. For example, in the skill *planning investigations*, the students analyzed the answering options more in depth focusing on the variables considered in the investigation, which reflects that they applied their epistemic knowledge to work on the multiple-choice items (category *methodological understanding: planning investigation*). The application of these general and skill-specific codes are seen as evidence based on response processes for the validity of the test score interpretation as measures of SRC (H 1b; cf. Mathesius et al. 2018a, b). In contrast, and especially interesting for the present study, a few verbalizations could be identified as indicating a lack of text comprehension or culture-related misunderstanding of the multiple-choice items (category *lack of text comprehension/indication of culture-related misunderstanding*). Within this category, three issues hint to problems resulting from the translation process, as described above. Roth et al. (2013) analyzed sources of DIF in science achievement tests and proposed (1) the relative text length, (2) linguistic issues, (3) the logical structure of an item (content or form), (4) cognitive-conceptual content, and (5) diversity issues. Two issues identified in the present study are linguistic issues in the sense of Roth et al. (2013): The use of the scientific term "carapace" in the English version of item "changing 2" instead of the direct translation "back shell" (*Rückenpanzer* in German) was due to the recommendation of the translators who judged the trivial term back shell as less commonly used in the English language. However, the findings of this study propose to add the trivial term in parentheses to each occurrence to make the item better understandable. The second linguistic issue identified in the present study is the ambiguous meaning of "might form the basis for this investigation," which is part of the four items "question 1", "question 2", "questions 3", and "hypotheses 2". As illustrated above, one pre-service teacher understood it as something which has to be done *prior* to the investigation. Hence, the formulation should be adjusted, for example, in "underlie this investigation", to make these items easier to understand. One issue identified in the present study relates to the logical structure of an item as discussed by Roth et al. (2013). The use of pronouns in the items "hypothesis 3" and "conclusion 2" made it difficult for the pre-service teachers to identify the logical relation between entities. This should be adjusted to warrant test equivalence, especially since the nouns (e.g., *Raupen* in German) are used repeatedly in the German items without replacing them by pronouns (Ercikan and Lyons-Thomas 2013; Roth et al. 2013).

The analysis of the think-aloud protocols provides possible sources of DIF. However, taking into account the empirical findings (Table 4), those items with minor or considerable DIF have not systematically been identified as problematic in the qualitative analysis. One exception is item "hypothesis 2" which showed considerable DIF which might have been caused by the ambiguous meaning of "might form the basis" as discussed above. However, this is unlikely because the other items including this formulation ("question 1", "question 2", "question 3") have not been identified as having problematic DIF. Hence, the findings of the qualitative analysis do not provide clear explanations for the observed DIF. Of course, this might be due to the rather small sample of pre-service teachers participating in the interviews ($n = 5$) resulting in only two to four qualitative analyses for each item. Another possible explanation might be specific emphases in teacher training in Germany and Australia, respectively, resulting in specific competencies of pre-service teachers in the two samples compared. The three items related to the skill *generating hypotheses* appeared to be more difficult in the English version than in the German version (MH statistic < 0; Table 4). Opposed to this, albeit not significant, all items related to *planning investigations* turned out to be easier in the English version (MH statistic > 0). As discussed by many authors, repeated and diverse scientific learning opportunities and explicit reflections about science and its procedures contribute to the development of SRC (Hartmann et al.

2015; Hodson 2014; Krell et al. 2015a; Mathesius et al. 2016). From this perspective, the findings may be a hint that the participating Australian pre-service science teachers possess lower skills related to the generation of hypotheses but higher skills to plan investigations than the participating German pre-service science teachers. This interpretation implies to further analyze and compare the teacher training curricula of the participating universities and identify areas of possible improvement.

As sketched out above, Australian pre-service science teachers are asked to develop SRC during teacher training (e.g., Won et al. 2017). H 2 examined to what extent the participating pre-service teachers are able to answer the present multiple-choice items adequately and, thus, provide evidence to possess an adequate level of SRC. The significant differences in the test scores between the pre-service secondary school teachers and the pre-service primary school teachers as well as the pre-service early childhood teachers can be interpreted as validity evidence based on relation to other variables (*known groups comparison*; AERA et al. 2014; Hartmann et al. 2015) since the pre-service secondary school teachers have had more opportunities to learn about science during their studies and, therefore, are expected to reach higher test scores (Hartmann et al. 2015; Mathesius et al. 2016).

Based on the equation of the 1PLM, three levels of SRC have been defined (basic, transitional, advanced; Table 7). The findings demonstrate that only about 13% of the entire Australian sample had advanced SRC, whereas 40% showed basic competencies; these findings illustrate the need to more explicitly emphasize on scientific reasoning in Australian teacher training.

One clear limitation of this study is the rather small and limited sample of 105 pre-service science teachers from one graduate school of education. However, since the focus of this study lies on test evaluation, this sample seems to be appropriate. Furthermore, additional sources of validity evidence may be obtained for the English-language instrument (cf. AERA et al. 2014; Kane 2013). Nevertheless, further studies may use the instrument presented here and evaluate to what extent the findings about SRC of pre-service science teachers in Australia are generalizable or an artifact of the sample analyzed here. Additional applications might use the instrument in order to obtain (convergent and discriminant) validity evidences of instruments assessing similar constructs or to conduct longitudinal analyses throughout the course of teacher education programs (but also beyond) of the development of SRC of English-speaking pre-service science teachers' SRC as it is done in other countries (Ding et al. 2016; Mathesius et al. 2016). The latter may contribute to detect effective opportunities to learn or to specifically improve teacher training in single study stages. Finally, the English, German, and Spanish versions of the instrument are currently used to assess SRC of pre-service teachers in different countries, in order to identify specific strength, weaknesses, and priorities in the science teacher education programs at the participating universities and enable specific improvements to what is taught to science teachers about scientific reasoning. More broadly, this will address international needs in science teacher education related to the assessment and development of pre-service teachers' SRC (cf. Osborne 2013).

Acknowledgments The authors thank the German Federal Ministry of Education and Research for funding the projects Ko-WADiS/ValiDiS (grant numbers 01PK11004A/01PK15004A) and the Center for International Cooperation of Freie Universität Berlin for funding the research in Melbourne.

Appendix

Table 8 Category system used for data analysis in this study (based on Mathesius et al. 2018b)

Category	Description
General categories	
Other	Not meaningful statements (e.g., requests about elapsed time or which item to answer next)
Reading/paraphrasing the task	Reading of the text or paraphrasing the content in own words
Lack of text comprehension/indication of culture-related misunderstanding	Verbalizations reflect a lack of text comprehension, a misunderstanding of the given information
Reasoned choice of the attractor	*The attractor is chosen* and reasons for the choice are provided (e.g., by paraphrasing the attractor to verify it or by further explaining the scientific approach described in the attractor)
Selection/exclusion based on signal word	*An answering option is chosen/excluded* because of a signal word (e.g., a word which is part of the item stem)
Selection/exclusion without a reason	*An answering option is chosen/excluded* without providing a reason related to the scientific problem; it may be argued that the answering option seems to be logical or attractive, or the decision may be result of a process of elimination
Categories for the skills *formulating questions* and *generating hypotheses*	
Alternative scientific approach	*A distractor is chosen* because it describes an alternative scientific approach which is seen as more important
Irrelevant factor	*An answering option is excluded* because the described scientific question/hypothesis is seen as irrelevant for the given scientific problem
Experimental testability impossible	*An answering option is excluded* because the described question/hypothesis is seen as scientifically not testable
Methodological understanding: scientific question/hypothesis	*An answering option is chosen/excluded* and the provided reasons reflect methodological understanding of scientific question/hypothesis
Methodological understanding missing: scientific question/hypothesis	*An answering option is chosen/excluded* and the provided reasons reflect a lack of methodological understanding of scientific question/hypothesis
Categories for the skill *planning investigations*	
Irrelevant design	*An answering option is excluded* because the described experimental design is seen as irrelevant for the given scientific problem
Methodological understanding: planning investigation	*An answering option is chosen/excluded* and the provided reasons reflect methodological understanding of planning investigations (e.g., principle "vary-one-thing-at-a-time")
Methodological understanding missing: planning investigation	
Categories for the skill *analyzing data and drawing conclusions*	
Criticism of method	*A distractor is chosen* because it describes criticism of the experimental design without proposing a conclusion
Inappropriate interpretation	*An answering option is excluded* because the described conclusion is seen as inappropriate within the given scientific problem (e.g., relation between variables, proposed causality)
Methodological understanding: analyzing data/drawing conclusion	*An answering option is chosen/excluded* and the provided reasons reflect methodological understanding of analyzing data/drawing conclusions (e.g., principle "vary-one-thing-at-a-time")

Table 8 (continued)

Category	Description
Categories for the skill *judging the purpose of models*	
Relevance of described purpose	*A distractor is chosen* because the described purpose is seen as interesting or principally relevant
Scientific application of models is denied	*An answering option is excluded* because the scientific application of models as research tools is principally denied
Criticism of described purpose	*An answering option is excluded* because the described purpose is seen as not appropriate (either principally or in within the given scientific problem)
Categories for the skill *testing models*	
Alternative scientific approach	*A distractor is chosen* but the provided reasons reflect a general understanding of the scientific application of models as research tools
Testing the model object	*An answering option is excluded* because it describes a test of the model object only
No empirical test	*An answering option is excluded* because it does not describe an empirical test of the model (either principally or in within the given scientific problem)
Criticism of model	*An answering option is chosen/excluded* and the provided reasons reflect criticisms of the model introduced in the item stem (e.g., too simplified for scientific purposes)
Categories for the skill *changing models*	
Alternative scientific approach	*A distractor is chosen* but the provided reasons reflect a general understanding of the scientific application of models as research tools
Model empirically falsified	*A distractor is chosen* and the provided reasons reflect the idea that a model is falsified by a single datum (i.e., should not be idealized)
Model not empirically falsified	*An answering option is excluded* and the provided reasons reflect the idea that a model is not falsified by a single datum (i.e., is idealized)
Inappropriate empirical basis	*An answering option is excluded* because it describes no or an insufficient empirical basis for changing the model

References

ACARA [Australian Curriculum, Assessment, and Reporting Authority] (2013). *General capabilities. January 2013 Edition*. Retrieved from http://docs.acara.edu.au/resources/General_Capabilities_2011.pdf.

AERA, APA, & NCME [American Educational Research Association, American Psychological Association, & National Council on Measurement in Education]. (2014). *Standards for educational and psychological testing*. Washington, DC: American Educational Research Association.

AITSL [Australian Institute for Teaching and School Leadership]. (2011). *Australian Professional Standards for Teachers*. Carlton South: Education Council Retrieved from https://www.aitsl.edu.au/docs/default-source/general/australian-professional-standands-for-teachers-20171006.pdf.

ASTA [Australian Science Teacher Association]. (2009). *National professional standards for highly accomplished teachers of science: Final draft*. Deakin: ASTA.

Baumert, J., & Kunter, M. (2013). The COACTIV model of teachers' professional competence. In M. Kunter, J. Baumert, W. Blum, U. Klusmann, S. Krauss, & M. Neubrand (Eds.), *Cognitive activation in the mathematics classroom and professional competence of teachers* (pp. 25–48). Boston: Springer US.

Bond, T., & Fox, C. (2001). *Applying the Rasch model: Fundamental measurement in the human sciences*. Mahwah: Erlbaum.

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

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

Capps, D., & Crawford, B. (2013). Inquiry-based professional development: What does it take to support teachers in learning about inquiry and nature of science? *International Journal of Science Education, 35*(12), 1947–1978. https://doi.org/10.1080/09500693.2012.760209.

Ding, L., Wei, X., & Mollohan, K. (2016). Does higher education improve student scientific reasoning skills? *International Journal of Science and Mathematics Education, 14*, 619–634. https://doi.org/10.1007/s10763-014-9597-y.

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

Educational Policies Commission. (1966). *Education and the spirit of science*. Washington, DC: National Education Association.

Embretson, S., & Reise, S. (2000). *Item response theory for psychologists*. Mahwah, NJ: Erlbaum.

Ercikan, K., & Lyons-Thomas, J. (2013). Adapting tests for use in other languages and cultures. In K. Geisinger (Ed.), *APA handbook of testing and assessment in psychology. Testing and assessment in school psychology and education* (pp. 545–569). Washington, DC: American Psychological Association.

Ercikan, K., Gierl, M., McCreith, T., Puhan, G., & Koh, K. (2004). Comparability of bilingual versions of assessments: Sources of incomparability of English and French versions of Canada's National Achievement Tests. *Applied Measurement in Education, 17*, 301–321. https://doi.org/10.1207/s15324818ame1703_4.

Ercikan, K., Arim, R., Law, D., Domene, J., Gagnon, F., & Lacroix, S. (2010). Application of think aloud protocols for examining and confirming sources of differential item functioning identified by expert reviews. *Educational Measurement: Issues and Practice, 29*, 24–35. https://doi.org/10.1111/j.1745-3992.2010.00173.x.

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

European Commission. (2015). *Science education for responsible citizenship*. Brussels: European Commission Retrieved from http://ec.europa.eu/research/swafs/pdf/pub_science_education/KI-NA-26-893-EN-N.pdf.

Forsyth, B., Kudela, M., Levin, K., Lawrence, D., & Willis, G. (2016). Methods for translating an English-language survey questionnaire on tobacco use into Mandarin, Cantonese, Korean, and Vietnamese. *Field Methods, 19*, 264–283. https://doi.org/10.1177/1525822X07302105.

Frey, A. (2006). Strukturierung und Methoden zur Erfassung von Kompetenz (Structuring and methods for competence assessment). *Bildung und Erziehung, 59*, 125–166.

Großschedl, J., Harms, U., Kleickmann, T., & Glowinski, I. (2015). Preservice biology teachers' professional knowledge: Structure and learning opportunities. *Journal of Science Teacher Education, 26*(3), 291–318. https://doi.org/10.1007/s10972-015-9423-6.

Hanushek, E., & Woessmann, L. (2011). How much do educational outcomes matter in OECD countries? *Economic Policy, 26*, 427–491. https://doi.org/10.1111/j.1468-0327.2011.00265.x.

Harkness, J. (2003). Questionnaire translation. In J. Harkness, F. J. R. van de Vijver, & P. Mohler (Eds.), *Cross-cultural survey methods* (pp. 35–56). Hoboken: Wiley.

Harkness, J., Pennell, B.-E., & Schoua-Glusberg, A. (2004). Survey questionnaire translation and assessment. In S. Presser, J. Rothgeb, M. Couper, J. Lessler, E. Martin, J. Martin, & E. Singer (Eds.), *Methods for testing and evaluating survey questionnaires* (pp. 453–473). Hoboken: Wiley.

Hartmann, S., Upmeier zu Belzen, A., Krüger, D., & Pant, H. (2015). Scientific reasoning in higher education. *Zeitschrift für Psychologie, 223*, 47–53. https://doi.org/10.1027/2151-2604/a000199.

Heijnes, D., van Joolingen, W., & Leenaars, F. (2017). Stimulating scientific reasoning with drawing-based modeling. *Journal of Science Education and Technology, 333*, 1096. https://doi.org/10.1007/s10956-017-9707-z.

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

Justi, R., & van Driel, J. (2005). A case study of the development of a beginning chemistry teacher's knowledge about models and modelling. *Research in Science Education, 35*, 197–219. https://doi.org/10.1007/s11165-004-7583-z.

Kane, M. (2013). Validating the interpretations and uses of test scores. *Journal of Educational Measurement, 50*, 1–73.

Kind, P., & Osborne, J. (2017). Styles of scientific reasoning: A cultural rationale for science education? *Science Education, 101*, 8–31. https://doi.org/10.1002/sce.21251.

Kleickmann, T., & Anders, Y. (2013). Learning at university. In M. Kunter, J. Baumert, W. Blum, U. Klusmann, S. Krauss, & M. Neubrand (Eds.), *Cognitive activation in the mathematics classroom and professional competence of teachers* (pp. 321–332). Boston: Springer US.

Klieme, E., Hartig, J., & Rauch, D. (2008). The concept of competence in educational contexts. In J. Hartig, E. Klieme, & D. Leutner (Eds.), *Assessment of competencies in educational contexts* (pp. 3–22). Göttingen: Hogrefe.

KMK (Ed.). (2017). *Ländergemeinsame inhaltliche Anforderungen für die Fachwissenschaften und Fachdidaktiken in der Lehrerbildung* (Common guidelines for the subjects and the subject didactics in teacher education). Berlin. Retrieved from https://www.kmk.org/fileadmin/Dateien/veroeffentlichungen_beschluesse/2008/2008_10_16-Fachprofile-Lehrerbildung.pdf.

KMK [Sekretariat der Ständigen Konferenz der Kultusminister der Länder in der BRD]. (2005). *Bildungsstandards im Fach Biologie für den Mittleren Schulabschluss (Biology education standards for the Mittlere Schulabschluss)*. München: Wolters Kluwer.

Krell, M., & Krüger, D. (2015). Testing models: A key aspect to promote teaching activities related to models and modelling in biology lessons? *Journal of Biological Education, 50*, 160–173. https://doi.org/10.1080/00219266.2015.1028570.

Krell, M., Koska, J., Penning, F., & Krüger, D. (2015a). Fostering pre-service teachers' views about nature of science: Evaluation of a new STEM curriculum. *Research in Science & Technological Education, 33*(3), 344–365. https://doi.org/10.1080/02635143.2015.1060411.

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

Krell, M. (2017). Schwierigkeitserzeugende Aufgabenmerkmale bei Multiple-Choice-Aufgaben zur Experimentierkompetenz im Biologieunterricht: Eine Replikationsstudie [Difficulty generating task characteristics of multiple-choice-tasks to assess experimental competencies]. *Zeitschrift für Didaktik der Naturwissenschaften.* https://doi.org/10.1007/s40573-017-0069-0.

Krell, M., Walzer, C., Hergert, S., & Krüger, D. (2017). Development and Application of a Category System to Describe Pre-Service Science Teachers' Activities in the Process of Scientific Modelling. *Research in Science Education, 333*, 1096. https://doi.org/10.1007/s11165-017-9657-8.

Krell, M., Vergara, C., van Driel, J., Upmeier zu Belzen, A., & Krüger, D. (2018). *Assessing pre-service teachers' scientific reasoning competencies: translation of a German mc instrument into Spanish/ English.* Paper presented at NARST conference 2018. USA: Atlanta, GA.

Kunter, M., Klusmann, U., Baumert, J., Richter, D., Voss, T., & Hachfeld, A. (2013). Professional competence of teachers: Effects on instructional quality and student development. *Journal of Educational Psychology, 105*, 805–820. https://doi.org/10.1037/a0032583.

Lawson, A. (2004). The nature and development of scientific reasoning: A synthetic view. *International Journal of Science and Mathematics Education, 2*, 307–338. https://doi.org/10.1007/s10763-004-3224-2.

Mathesius, S., Upmeier zu Belzen, A., & Krüger, D. (2014). Kompetenzen von Biologiestudierenden im Bereich der naturwissenschaftlichen Erkenntnisgewinnung: Entwicklung eines Testinstruments [Competencies of biology students in the field of scientific inquiry: Development of a testing instrument]. *Erkenntnisweg Biologiedidaktik, 13*, 73–88.

Mathesius, S., Hartmann, S., Upmeier zu Belzen, A., & Krüger, D. (2016). Scientific reasoning as an aspect of pre-service biology teacher education. In T. Tal & A. Yarden (Eds.), *The future of biology education research*. Proceedings of the 10th conference of European Researchers in Didactics of Biology (ERIDOB) (pp. 93–110). Haifa, Israel.

Mathesius, S., Upmeier zu Belzen, A. & Krüger, D. (2018a). Eyetracking als Methode zur Untersuchung von Lösungsprozessen bei Multiple-Choice-Aufgaben zum wissenschaftlichen Denken. In: M. Hammann & M. Lindner (Hrsg.), *Lehr- und Lernforschung in der Biologiedidaktik, Band 8* (pp. 225–244). Innsbruck: Studienverlag.

Mathesius, S., Upmeier zu Belzen, A. & Krüger, D. (2018b). *Lautes Denken bei der Bearbeitung von Multiple Choice Aufgaben zur Erfassung von Kompetenzen des wissenschaftlichen Denkens (working title).* Manuscript in preparation.

Mayer, J. (2007). Erkenntnisgewinnung als wissenschaftliches Problemlösen (Scientific inquiry as problem solving). In D. Krüger & H. Vogt (Eds.), *Theorien in der biologiedidaktischen Forschung* (pp. 177–186). Berlin: Springer.

Mayer, D., Sodian, B., Koerber, S., & Schwippert, K. (2014). Scientific reasoning in elementary school children: Assessment and relations with cognitive abilities. *Learning and Instruction, 29*, 43–55. https://doi.org/10.1016/j.learninstruc.2013.07.005.

Morris, B., Croker, S., Masnick, A., & Zimmerman, C. (2012). The emergence of scientific reasoning. In H. Kloos, B. Morris, & J. Amaral (Eds.), *Current topics in children's learning and cognition* (pp. 61–82). InTech.

Neumann, K., Härtig, H., Harms, U., & Parchmann, I. (2017). Science teacher preparation in Germany. In J. Pedersen, T. Isozaki, & T. Hirano (Eds.), *Model science teacher preparation programs. An international comparison of what works* (pp. 29–52). Information Age: Charlotte.

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

OECD. (2010). *The high cost of low educational performance: The long-run economic impact of improving PISA outcomes*. Paris. Retrieved from https://www.oecd.org/pisa/44417824.pdf.

Opitz, A., Heene, M., & Fischer, F. (2017). Measuring scientific reasoning: A review of test instruments. *Educational Research and Evaluation, 23*, 78–101. https://doi.org/10.1080/13803611.2017.1338586.

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

Osborne, J. (2014). Scientific practices and inquiry in the science classroom. In N. Lederman & S. Abell (Eds.), *Handbook of research on science education* (pp. 579–599). New York: Routledge.

Pedersen, J. E., Isozaki, T., & Hirano, T. (Eds.). (2017). *Model science teacher preparation programs: An international comparison of what works*. Charlotte: Information Age.

Roth, W.-M., Oliveri, M., Sandilands, D., Lyons-Thomas, J., & Ercikan, K. (2013). Investigating linguistic sources of differential item functioning using expert think-aloud protocols in science achievement tests. *International Journal of Science Education, 35*, 546–576. https://doi.org/10.1080/09500693.2012.721572.

Schauble, L., Klopfer, L., & Raghavan, K. (1991). Students' transition from an engineering model to a science model of experimentation. *Journal of Research in Science Teaching, 28*, 859–882.

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

Schreier, M. (2012). *Qualitative content analysis in practice*. Thousand Oaks: Sage.

Schwarz, C., & White, B. (2005). Metamodeling knowledge: Developing students' understanding of scientific modeling. *Cognition and Instruction, 23*, 165–205.

Shulman, L. (1986). Those who understand: Knowledge growth in teaching. *Educational Researcher, 15*, 4–14.

Stiller, J., Hartmann, S., Mathesius, S., Straube, P., Tiemann, R., Nordmeier, V., … Upmeier zu Belzen, A. (2016). Assessing scientific reasoning: A comprehensive evaluation of item features that affect item difficulty. Assessment & Evaluation in Higher Education, 41, 721–732. doi:https://doi.org/10.1080/02602938.2016.1164830

Thompson, E., Bowling, B., & Markle, R. (2017). Predicting student success in a major's introductory biology course via logistic regression analysis of scientific reasoning ability and mathematics scores. *Research in Science Education, 30*(2), 663–163. https://doi.org/10.1007/s11165-016-9563-5.

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

van der Graaf, J., Segers, E., & Verhoeven, L. (2016). Scientific reasoning in kindergarten: Cognitive factors in experimentation and evidence evaluation. *Learning and Individual Differences, 49*, 190–200. https://doi.org/10.1016/j.lindif.2016.06.006.

VCAA [Victorian Curriculum and Assessment Authority]. (2016a). *Victorian certificate of education biology: Advice for teachers*. Melbourne: VCAA.

VCAA [Victorian Curriculum and Assessment Authority]. (2016b). *Victorian Curriculum: F-10.* Melbourne, VIC. Retrieved from http://victoriancurriculum.vcaa.vic.edu.au/science/curriculum/f-10.

Weinert, F. (2001). Concept of competence: A conceptual clarification. In D. Rychen & L. Salganik (Eds.), *Defining and selecting key competencies* (pp. 45–65). Kirkland: Hogrefe.

White, B., Collins, A., & Frederiksen, J. (2011). The nature of scientific meta-knowledge. In M. Khine & I. Saleh (Eds.), *Models and modeling. Cognitive tools for scientific enquiry* (pp. 41–76). Dordrecht: Springer.

Windschitl, M., Thompson, J., & Braaten, M. (2008). Beyond the scientific method: Model-based inquiry as a new paradigm of preference for school science investigations. *Science Education, 92*(5), 941–967. https://doi.org/10.1002/sce.20259.

Won, M., Hackling, M., & Treagust, D. (2017). Secondary science teacher education in Australia. In J. Pedersen, T. Isozaki, & T. Hirano (Eds.), *Model science teacher preparation programs. An international comparison of what works* (pp. 229–248). Information Age: Charlotte.

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

Zwick, R., Thayer, D., & Lewis, C. (1999). An empirical Bayes approach to Mantel-Haenszel DIF analysis. *Journal of Educational Measurement, 36*, 1–28.