Advancing the Assessment of Scientific Reasoning Skills: A Review of Tests and a Detailed Analysis of a Common Test

Dissertation zum Erwerb des Doctor of Philosophy (Ph.D.)

am Munich Center of the Learning Sciences

der Ludwig-Maximilians-Universität

München

Vorgelegt von

Ansgar Opitz

10.11.2016

1st Supervisor / Erstgutachter:	Prof. Dr. Moritz Heene
2nd Supervisor / Zweitgutachter:	Prof. Dr. Frank Fischer
Tag der mündlichen Prüfung:	30.1.2017

Acknowledgements

The research presented in this work was supported by the Elite Network of Bavaria [Project number: K-GS-2012-209]. I would like to extend my sincere gratitude to the opportunities (conference attendance, incubator stay, international knowledge exchange) made possible by the ENB.

Study 1 was submitted for publication and (slightly edited) published as:
Opitz, A., Heene, M., & Fischer, F. (2017). Measuring scientific reasoning – a review of test instruments. *Educational Research and Evaluation*, *23*(3–4), 78–101. https://doi.org/10.1080/13803611.2017.1338586

A publication of Study 2 is planned but no journal has been determined for submission yet.

The underlying work for and the completion of this thesis would not have been possible without the support of a number of people. In the next paragraphs I want to express my immense gratitude to these people.

First, I want to thank my supervisors. I want to thank Professor Heene for his impeccable guidance and wisdom regarding the statistics and research methods of my thesis, and in general for sharing my enthusiasm regarding open science all things statistical. I want to thank Professor Fischer for not letting me get away with easy answers and for his orientation in the jungle of the research world. I want to thank both of them for our many productive meetings, as well as their trust and their confidence in my scientific reasoning skills. Additionally, I want thank Professor Richard Shavelson and Professor Jonathan Osborne for invaluable advice and their hospitality when I visited Stanford. Thank you!

Second, I would like to thank my mom and my dad, and my siblings, Birte and Thoren, for the great effect ($d = 0.8$) they had on me since – forever, really. Thank you!

Third, I want to thank my Reason colleagues. They truly are an 11-factor solution for a great intercultural and diverse work experience. I want to thank you for all the shared meals, the dance recesses, and most of all, for teaching me so many things both about research and the world in general! In particular, I want to thank the best office buddies in the world ($p < .001$), Janina, Maryam, and Sandra, for all the (non)sense you brought to my office life. Thank you!

Fourth, I want to thank several friends, whom I trust more than I would trust a BF > 100. I want to thank the members of the Psycho-Stammtisch, Esther, Hannah, Hanni, Janina, Kat, Lena, Nici, and Sophie who manage to accomplish the task that was thought to be impossible of making Mondays (and also every other moment you spend with them) absolutely awesome. I also want to thank my other friends in Munich and around the world who supported me on my PhD journey, especially Maxi, Sarah, Andy, Steffi, Julia, Angi, Eline, Eiko, Paulette, Taylor, Cameron, Chris, Greg, Julia, Alice, Lucie, and Stacey. All of you explained most of the variance of my happiness in the last years and – as a great motivational poster once said – always managed to brighten the gloomiest of days (of which there were only a few thanks to you). Expressing the true amount of my gratitude towards you is as impossible as proving the H_0. Let us just say, that I feel greatly indebted to you since, as you can see in this acknowledgement, you *almost* managed the colossal task of keeping me from going insane. Thank you!

Fifth, a special thank you to those of you who supported me with proofreading, feedback, mental support, and in a thousand other small ways during these last months! Thank you!

Last, thanks to whoever invented peanut butter cookies. I might still be looking for participants if not for these great treats. Thank you!

Table of Contents

Abstract

Acquiring scientific reasoning skills, such as the construction of hypotheses or the generation and evaluation of evidence, is considered an important educational goal. However, much is unknown about the construct of scientific reasoning: Which skills make up the construct? Is there a single scientific reasoning dimension or multiple dimensions? Is scientific reasoning domain-general or domain-specific? Is it distinct from general reasoning? Similarly, there are knowledge gaps regarding the assessment of scientific reasoning: Which tests exist? How do they conceptualize scientific reasoning? What are their psychometric properties? Two studies were conducted to investigate these open questions about the scientific reasoning construct and its assessment. The first study reviewed 38 scientific reasoning tests. An analysis of the skills that were assessed showed that the tests focused on *hypothesis generation*, *evidence generation*, *evidence evaluation*, and *drawing conclusions*. Additionally, newer tests recognized a more diverse set of kills, but they did not increase the total number of assessed skills. Over time, conceptualizations of scientific reasoning have moved towards a domain-specific set of different but coordinated skills and away from domain generality and unidimensionality assumptions. Test authors rarely tested their assumptions and controls of psychometric properties were lacking. In the second study, a scientific reasoning test and three general reasoning tests were given to 507 university students from physics, biology, and medicine. Although physics students achieved the highest average scores on the scientific reasoning test, differential item functioning analyses revealed that several items with a biology context where biased in favor of biology students. While regression analyses showed

a moderate overlap between scientific and general reasoning, a bifactor model analysis indicated that the scientific reasoning test contained hardly any variance that was not accounted for by general reasoning. Items from the scientific reasoning test that had a strong quantitative reasoning component loaded negatively on a scientific reasoning factor in the model. An exploratory factor analysis showed that both a one-factor and a seven-factor model fit the data well. Taking the results from both studies into account, the following conclusions can be drawn: First, while it is too early to settle on an exact set of skills that makes up scientific reasoning, it seems as if conceptualizations that do not include quantitative reasoning are better suited for the construction of valid tests. Second, scientific reasoning is neither completely domain-general nor domain-specific and a compromising position seems most promising. Third, the overlap between scientific and general reasoning might be higher than previously thought. Fourth, the structure of scientific reasoning is complicated and seems to contain a weak general factor and several minor factors. All of these aspects, along with other checks of psychometric properties, should be investigated further in the future and the procedures suggested in the second study, as well as procedures from other fields of assessment, offer possibilities to do so. Additionally, a range of scientific reasoning tests exists and the conceptualizations of scientific reasoning that are used to construct them become more nuanced, i.e. they recognize a more diverse set of skills and move away from unidimensionality and domain generality assumptions. However, high-stake decisions should not be based on scientific reasoning tests at the moment given the remaining open questions regarding both the construct of scientific reasoning and the psychometric properties of its assessments.

1 General Introduction

The assessment of scientific reasoning has been called the 21st century challenge for science education (Osborne, 2013). Scientific reasoning skills entail a range of capacities. A person possessing good scientific reasoning skills will be able to identify relevant problems as well as questions and to produce high quality hypotheses, prototypical artefacts, and evidence. Furthermore, they will succeed in using evidence in order to draw meaningful conclusions about a claim and in communicating with others about the results of a scientific reasoning process. What is so special about scientific reasoning skills, which are included in almost every conceptualization of the broader construct of scientific literacy (Norris, Phillips, & Burns, 2014), that their assessment is of such importance?

Several arguments for the importance of scientific literacy and thus also scientific reasoning skills can be found in the literature: They are useful on a personal level as an underlying skill set when engaging in inquiry learning (Lazonder & Harmsen, 2016). Mastering scientific reasoning skills, such as formulating questions, evaluating evidence, and using this evidence to judge a claim, is also in the interest of the general public as these skills are a critical component for civic engagement (Rudolph & Horibe, 2016). This was emphasized by a collection of case studies showing that skills such as evaluating and interpreting data are not only important according to abstract policy guidelines but also for people when they actually encounter science in their daily lives (Ryder, 2001). Furthermore, Mc Eneaney (2003) assumed that the spread of scientific literacy is in the interest of politicians to keep up the support of science. Another argument in favor of the critical value

of scientific reasoning skills that has been made for decades is economic. According to this view scientific reasoning is a vital part of the skill set of a productive workforce, which in turn is good for the economy (Osborne, 2013; The Royal Society, 2014). Additionally, the literature mentions a cultural argument, saying that scientific literacy is necessary to understand science as an important achievement of humanity (Ryder, 2001). These diverse arguments in favor of scientific literacy, and scientific reasoning as part of it, are not specific to a certain region of the world but instead common around the globe (Mc Eneaney, 2003).

Given all these arguments for the relevance of scientific reasoning skills it is not surprising that the assessment of these skills is considered as a major challenge for science education in the 21st century. This thesis contributes to meeting this challenge. It presents two studies that deal with the assessment of scientific reasoning skills. Study 1 is a review of scientific reasoning tests and Study 2 analyzed a common scientific reasoning test in great detail. Before the two studies are presented, this Introduction will provide important background information. It begins by describing some of the most influential scientific reasoning conceptualizations of the last few decades. This is followed by a description of the scientific reasoning conceptualization that is used in this thesis and an explanation about why it is used. The conceptual clarifications continue by pointing out differences between scientific reasoning and neighboring constructs, defining *skill*, and providing an overview of different terms for scientific reasoning skills, and they finish with several minor but still relevant conceptual aspects. The next section will present the debate around the domain generality vs. domain specificity of scientific reasoning. It will be argued that this debate is not only an important theoretical debate but that it also has direct consequences for the assessment of scientific reasoning. The Introduction continues with an overview of psychometric properties that will play a central role in the presented studies and with an introduction to the Rasch model and differential item functioning. This section is

complemented by a brief history of scientific reasoning tests, demonstrating that all the aspects mentioned in the preceding sections are indeed relevant to the assessment of scientific reasoning using concrete examples. The Introduction closes with a summary of the research goals and questions and a brief overview of the two studies of this thesis. After the presentation of these two studies in the central part of this work, this thesis will conclude with a General Discussion that brings together the results of the two studies, addresses implications and limitations, and gives suggestions for future research.

1.1 The Conceptualization of Scientific Reasoning

1.1.1 Influential scientific reasoning conceptualizations

It is common in the social sciences for multiple conceptualizations to exist for the same construct, and scientific reasoning is no exception to this phenomenon. Before the conceptualization that will be used in this thesis is introduced, the next paragraphs will give a chronological overview of selected influential scientific reasoning conceptualizations and their characteristics that are most important for this thesis. This overview focuses on prime representatives of scientific reasoning conceptualizations. A more detailed analysis of conceptualizations used for test construction can be found in Study 1.

One conceptualization that had a big influence on the research about scientific reasoning is the theory about the stages of the development of human thinking by Inhelder and Piaget (1958). The highest stage according to this theory, formal operational reasoning, includes the evaluation of hypotheses using evidence, a skill which is widely recognized as a main scientific reasoning skill. The theory assumes that this skill is part of a developing general reasoning ability. Furthermore, it is assumed that this skill can be applied without domain restrictions. These two assumptions are not made by newer conceptualizations.

Additionally, the evaluation of hypotheses using evidence, as important as this skill certainly is, is seen as just one of several subparts of scientific reasoning by more recent conceptualizations.

One example for this is the second conceptualization that had a major influence on the field, the Scientific Discovery as Dual Search (SDDS) model (Klahr & Dunbar, 1988). In the SDDS model, scientific reasoning is described as a problem-solving process which already recognizes three components: A search in a hypothesis space in order to construct hypotheses, a search in an experiment space in order to generate evidence, and the subsequent analysis of the evidence. While these three skills are connected within a problem-solving process, they are not thought of as indistinguishable parts of one general scientific reasoning ability. This is demonstrated in one of the studies by Klahr and Dunbar (1988) when they asked their participants to begin the process by selectively focusing on the search in the hypothesis space. The authors also recognized the importance of domain-specific knowledge and thus set a limit to the domain generality of scientific reasoning skills. Problem-solving process approaches to scientific reasoning remained prominent during the following decades (J. Mayer, 2007; Zimmerman, 2000). Even in a recent conceptualization of scientific activity by Osborne (2013) the SDDS model makes up one of two major parts.

Despite their popularity, these older scientific reasoning conceptualizations were also criticized as being proponents of a universal scientific method, giving off the impression that there is only a single prevailing way to conduct research (H. H. Bauer, 1994). In educational institutions this universal method takes the form of exercises in which students are given a fixed narrow claim that is tested with an experiment (Windschitl, Thompson, & Braaten, 2008). Although the idea of such a standard scientific method was uncontested for a long time, alternatives have been suggested recently. For instance, Windschitl et al. (2008) suggested to teach scientific reasoning as model-based inquiry, which focuses on constructing

a model of a phenomenon using different types of evidence instead of only experiments. The critique of the universal scientific method was probably one reason why modern conceptualizations of scientific reasoning often contain more skills, and do not only focus on experiments as a method to gather data. One of the most prominent modern conceptualizations that exemplifies this trend is the most recent framework of the National Research Council (National Research Council [NRC], 2012). It distinguishes between eight scientific and engineering practices: *asking questions and defining problems*, *developing and using models*, *planning and conducting investigations*, *analyzing and interpreting data*, *using mathematics and computational thinking*, *constructing explanations and designing solutions*, *engaging in arguments from evidence*, and *obtaining, evaluating and communicating information*. They are applicable in four disciplinary domains: physical sciences, life sciences, earth and space sciences, and engineering.

1.1.2 A descriptive framework of scientific reasoning skills

The conceptualization that will be the basis throughout this review is similar to the conceptualization of the NRC (2012). It is a framework by F. Fischer et al. (2014) that also recognizes eight scientific reasoning skills. A description of the skills along with examples can be seen in Table 1. The next paragraphs will point out several aspects of this framework that deserve special attention.

The framework is descriptive in its nature and not normative. It is not implied that these are the only possible scientific reasoning skills one could think of, that this is the only possible conceptualization of scientific reasoning, or that students have to develop exactly these scientific reasoning skills. This is in line with the position of diSessa (2008): There are multiple interpretations of scientific reasoning and it might not be possible to decide which is the right one. Thus, it is more useful to describe where your research stands and how it relates

to the work of others. However, diSessa also emphasizes that we should try to find as much consensus as we can in the form of a small number of core models. The framework by F. Fischer et al. (2014) is a good candidate for such a core model. It is especially a good candidate for a core model that can serve as the basis for a review, which was the case in Study 1, for the following reasons.

Table 1

List of Scientific Reasoning Skills With Examples

Skill name	**Skill description**	**Example**
Problem identification	Perceiving a mismatch between a problem (from a science, professional, or real-world context) and current explanations, analysing the situation, and building a problem representation.	Filtering diagnostic information from a patient's description of their symptoms.
Questioning	Identifying one or more questions as the basis for an upcoming reasoning process.	Asking how we can determine if an Euler-Walk is possible.
Hypothesis generation	Constructing possible answers to a question (according to scientific standards) based on known models, frameworks, or evidence.	Evidence-based suggestion of a substance that influences how the memory of honey bees develops.
Construction and redesign of artefacts	Creating a prototypical artefact, testing it, and revising it based on the test.	Building a prototype of a learning environment based on a theory.
Evidence generation	Producing evidence following one of several methods. Amongst them are controlled experiments, observational studies, and deductive reasoning based on a theory.	Collect point-line configurations that allow or prohibit an Euler-Walk.
Evidence evaluation	Analysing various forms of evidence in regard to a claim or theory.	Evaluate possible diagnoses with evidence from physical examinations, lab tests, etc.
Drawing conclusions	Coming to a conclusion by weighing the relevance of different pieces of evidence. Can lead to the revision of an initial claim.	Using evidence to decide if a learning environment needs to be re-designed.
Communicating and scrutinizing	Presenting and discussing the methods and the results of a scientific reasoning process both within a team and a broader community.	Giving a talk about the results of a honey bee study.

Note. Adapted from F. Fischer et al. (2014).

The framework is very inclusive, so other frameworks can be related to it. For instance, all of the skills from the theory about the development of human thinking, the SDDS model, and most of the skills by the NRC (2012) find their expression in the framework. The one difference from the NRC (2012) model that seems noteworthy is that the framework by F. Fischer et al. (2014) does not contain quantitative reasoning. Quantitative reasoning is considered to be a higher level cognitive skill, instead (Shavelson & Huang, 2003). This difference highlights one crucial aspect of scientific reasoning that is still debated, namely which skills are part of scientific reasoning. While the older conceptualizations by Inhelder and Piaget (1958) and Klahr and Dunbar (1988) only recognized a subset of the skills in the conceptualization by F. Fischer et al. (2014), other conceptualizations, e.g. the one by the NRC (2012), added other skills to their set. This debate about which skills are part of scientific reasoning is also addressed by both studies in this thesis.

Another advantage of the framework by F. Fischer et al. (2014) over other scientific reasoning conceptualizations is that a variety of disciplines (psychology, education, biology, medicine, mathematics, media informatics, and social work) collaborated in its creation. Thus, it already demonstrated that it provides a shared language for researchers of scientific reasoning skills independent of their discipline. This is especially important because this thesis is not only looking at scientific reasoning in the natural sciences. Instead, scientific reasoning skills are also considered to be important in the social sciences, mathematics, medicine, etc., which can be seen from the examples in Table 1. Of course, the concrete form a skill takes in a discipline can differ and the skills can be of different importance. It should also be noted that the skills are not necessarily part of a sequential process. Some skills will probably often occur in a sequence (e.g. *evidence evaluation* following *evidence generation*) but it is also rare that all eight skills will occur in a set sequence without jumping back and forth or leaving out certain steps.

Lastly, the framework by F. Fischer et al. (2014) recognizes both domain-general and domain-specific aspects of scientific reasoning, so it is not biased towards one of these positions (a detailed introduction into this debate is given in a later section). Overall, this is why the framework was selected for this thesis: It strikes a balance between being inclusive enough not to miss important contributions to the assessment of scientific reasoning, while being exclusive of higher level reasoning skills. If we would include higher level reasoning skills such as quantitative reasoning in our conceptualization, and consequently consider all the according assessments, the scope of this work would become so broad that it would be difficult to draw conclusions about a distinct field of research.

1.1.3 Differences between scientific reasoning, nature of science, and scientific argumentation

When one defines a framework of what scientific reasoning skills are, it is also necessary to define what scientific reasoning skills are not. A first distinction is made between scientific reasoning skills and the understanding of the nature of science (NOS). This distinction seems especially relevant since these two concepts often appear together in frameworks for scientific competencies or scientific literacy (Holbrook & Rannikmae, 2009; J. Mayer, 2007). In comparison to scientific reasoning, there is a wide consensus about what belongs to NOS (Osborne, Collins, Ratcliffe, Millar, & Duschl, 2003). NOS has these major aspects: Knowledge is tentative, empirically based, subjective, partly a product of imagination and creativity, and socially and culturally embedded (Khishfe, 2008). These aspects are also reflected in common assessments of NOS such as the Views of Nature of Science Questionnaire (VNOS; Lederman, Abd-El-Khalick, Bell, & Schwartz, 2002) and others (Liang et al., 2008).

Now, what is the difference between scientific reasoning and NOS? NOS is normative and not the same as scientific practice (Abd-El-Khalick, 2012). Phrased differently, NOS is more about understanding scientific reasoning skills from an epistemic perspective and not about using them. It should be noted that while this distinction is clear in theory, it is not always as easy to distinguish the two concepts in practice. This is especially the case with the research about practical epistemology (Sandoval, 2005). The focus of this research field is to observe students' beliefs about knowledge construction, which is closely related to NOS, while they engage in scientific inquiry, which requires scientific reasoning skills. Similar overlaps can occur in tests. For instance, a test item could be about finding flaws in an experiment. In such a situation it is important to explore the intention of the test's author. Is the test asking this question because it wants to test the ability to construct a correct experiment (which indicates a closer relation to scientific reasoning) or does it aim at understanding the difficulties in generating scientific knowledge (which indicates a closer relation with NOS)? To summarize, scientific reasoning and NOS can overlap in practice but it is important to note that they are different on a conceptual level.

The second distinction that should be made is between scientific reasoning and scientific argumentation. Research about scientific reasoning focuses on the cognitive skills of individuals, which are assessed individually. In comparison, argumentation research is focused on the social context and engaging in discussions with others (Berland & McNeill, 2010; Noroozi, Weinberger, Biemans, Mulder, & Chizari, 2012; Osborne, 2010; Weinberger, Stegmann, & Fischer, 2010). There is also a separate body of interventions for argumentation (Cavagnetto, 2010). As with NOS, the distinction is not always clear-cut. Discussions have their place in scientific reasoning frameworks but they are placed at the end or accompanying a scientific reasoning process and are not central to it (Pedaste et al., 2015).

1.1.4 A definition of "skill"

After describing how scientific reasoning is conceptualized, the use of the word *skill* also needs some explanation, because for some researchers scientific reasoning is just about knowledge. For instance, in a framework by Li, Ruiz-Primo, and Shavelson (2006) a task like graph analysis, which can be described as *evidence evaluation*, is categorized under procedural knowledge. However, this thesis is adopting a different approach. According to skill theory (K. W. Fischer, 1980) a skill is a behavior and it can be cognitive. Skills consist of organized sequences; a single piece of information is not enough to form a skill (Fitts & Posner, 1967). Similarly, this thesis takes the stance that scientific reasoning is not just pure knowledge but at least the application of knowledge. In a task like analyzing a graph, knowledge certainly plays a role, but it is either unconsciously used or actively retrieved. Such a process is more than just knowledge. It includes an action that is captured in the word *skill*. Again, when it comes to testing, the distinction between knowledge and skill can be difficult on a surface level. A test's scoring rubric might award points for understanding the effects of sample size on an experiment, which sounds a lot like knowledge, but when we look at the test itself it might be very clear that the intention is to measure the ability to construct an experiment, which is very much skill related (Sirum & Humburg, 2011).

Another very important feature of theories about skills is the possibility of learning new skills: Practice can lead to acquiring new skills or strengthening old ones (Anderson, 1982, 1993, 1996; Fitts & Posner, 1967). Similarly, when a skill is not used it will get weaker (Anderson & Schunn, 2000). The possibility of influencing a skill via training in a short period of time marks a crucial difference between the *skills* and the *traits* of a person, such as intelligence, which are assumed to be stable over long periods of time and will hardly react to training (Asendorpf, 2011; Cattell, 1971). This implies that scientific reasoning skills can be

trained and thus they are not meant be the same as intelligence or general reasoning[1]. Evidence from a meta-analysis supports the notion that scientific reasoning can be trained (Engelmann, Neuhaus, & Fischer, 2016).

1.1.5 Different terms for scientific reasoning skills

Another relevant aspect when dealing with scientific reasoning skills on a conceptual level is that the same skills are sometimes described using different terms. This is apparent when we look at four skills from the F. Fischer et al. (2014) conceptualization that can be found in many tests: *hypothesis generation*, *evidence generation*, *evidence evaluation*, and *drawing conclusions*. Tasks from tests that aimed at these four skills have been described with different terms in the literature. Apart from scientific reasoning one can find the terms *scientific thinking* (Azarpira et al., 2012), *science process skills* (Feyzioglu, Demirdag, Akyildiz, & Altun, 2012), *experimental problem-solving skills* (Ross & Maynes, 1983), *scientific literacy* (H.-P. Chang et al., 2011), *scientific inquiry* (Gobert, Sao Pedro, Raziuddin, & Baker, 2013) and *problem-solving* (Shaw, 1983). Additionally, *evidence generation* tasks were described as *science practices* (United States National Assessment Governing Board, WestEd (Organization), & Council of Chief State School Officers, 2010), *scientific competencies* (Organisation for Economic Co-operation and Development [OECD], 2006), *research skills* (Meerah et al., 2012), *procedural understanding* (Roberts & Gott, 2004), and *formal reasoning* (Tobin & Capie, 1981). However, this thesis is more interested in the skills themselves than the names by which they are described. The term *scientific reasoning* is deemed best because it is specific enough to indicate a difference from NOS and

[1] The term *general reasoning* will be used in this thesis as a collective term for any subset of intelligence facets, but mainly the combination of verbal, numerical, and figural reasoning that is tested in Study 2. Considering the way these three scales are assessed in Study 2, it would not be justified to describe them as a complete measure of intelligence. Thus, it seemed more appropriate to use the term *general reasoning* when talking about the connection between scientific reasoning and higher-order cognitive constructs.

argumentation but also broad enough to be interdisciplinary and inclusive of a variety of skills.

1.1.6 Further conceptual clarifications

The last part of this section covers several shorter clarifications regarding the concept of scientific reasoning. The first one is about the relation between the scientific reasoning of experts and the scientific reasoning of laypersons. On the one hand, there is the perspective that the scientific reasoning of students and the scientific reasoning of scientists is fundamentally different (Chinn, 2006). This is also implied when it is said that educating citizens and educating scientists is not the same (Bybee, McCrae, & Laurie, 2009). On the other hand, there is the perspective that a continuum exists between laypersons and scientists (Eraña & Martínez, 2004). This position is often adopted by developmental psychologists who point out that even young children engage in hypothesis testing using data (Gopnik, 2012; Xu & Kushnir, 2013) and that children are better at scientific reasoning than one would normally expect (Zimmerman, 2007).

A useful concept to bridge the difference between the two positions is the idea of epistemic modes (F. Fischer et al., 2014). According to this concept, people engaging in scientific reasoning can have different goals. Some are more interested in the advancement of theories while others have practical applications in mind. The underlying skills are the same in both cases, though. This thesis adopts the continuum approach. On this continuum, the presented studies will focus on learners of scientific reasoning and not on experts. Since learners of scientific reasoning are the main target group for trainings and interventions, the assessment of scientific reasoning is especially important in this group.

Related to this topic is the question whether a skill is only a scientific reasoning skill when it is applied in a science context. While some have defined scientific reasoning this way

(Harlen, 1999), this thesis is taking a different approach. As was stated when the framework for scientific reasoning was introduced, scientific reasoning is not only relevant in the natural sciences. This is in line with the position that skills like the coordination of evidence and theory are also relevant outside of science (Kuhn, 2002). A last clarification that should be made is that this thesis is about scientific reasoning skills that individuals use, or should use, to solve problems and not about approaches to solve scientific problems throughout the history of science, which is more aligned with the idea of styles of scientific reasoning (Kind & Osborne, 2016).

1.2 The Debate About the Domain Generality vs. Domain Specificity of Scientific Reasoning

One highly debated aspect of scientific reasoning is whether it is domain-general or domain-specific. Since so many voices have contributed to this debate this entire section is dedicated to it. After clarifying what a domain is, this section presents the different positions in this debate, giving arguments in favor of domain generality, in favor of domain specificity, and in favor of a compromising position. The studies presented in this thesis will contribute to this crucial theoretical debate but of course they will have a specific focus on how this debate is relevant to the assessment of scientific reasoning. Thus, at the end of this section, once the necessary understanding of the debate is established, the consequences of the debate for the assessment of scientific reasoning will be highlighted.

1.2.1 Different understandings of domain

Before we can get into the debate about whether scientific reasoning is domain-general or domain-specific we have to clarify the term *domain*. The term can refer to different levels:

First, the term can refer to a group of subject fields like the natural sciences or the social sciences (van Gigch, 2002). Second, it can refer to a specific subject field, such as physics, chemistry, biology, medicine, psychology, etc. (Erduran, 2007). Third, it can refer to one research area within one subject field, e.g. cognitive or social psychology (Schunn & Anderson, 1999).

In Studies 1 and 2 the main focus is on the distinction between different subject fields. The first reason for this focus is simply that test authors often align their test contexts and their assumptions about domain generality along different subject fields (Cloonan & Hutchinson, 2011; Germann, 1989; Grube, 2010; Hammann, Phan, Ehmer, & Grimm, 2008). The second reason is – and this is probably also one reason why many test authors choose this focus – that focusing on subject fields is useful for the operationalization of a study. It allows separations of people and tasks into groups that are closely aligned with different subjects taught at school and majors chosen at university. Consequences from test results would also very much influence decisions along these subject field lines, e.g. the possibility of assessing groups from different majors together. Of course, adopting this position does not mean that other definitions cannot also be useful in other circumstances.

1.2.2 Overview of different positions in the debate

In the following paragraphs, arguments for both sides of the debate as well as a compromising point of view are given. Before diving into the arguments, it needs some notes in advance: In order to not miss relevant contributions, the debate includes arguments for domain generality or domain specificity independent of the specific level that *domain* refers to in the source material. Domain generality implies, independent of the exact definition of domain, that there is a skill or multiple skills that function similarly in different contexts independent of domain-specific aspects like knowledge. Thus, a significant transfer of skills

is possible. Domain specificity means that scientific reasoning is so closely tied to a domain that there are no shared aspects of scientific reasoning among domains. Thus, the transfer of skills will be very limited or not possible at all.

Many studies that deal with this debate stem from the area of developmental psychology. This is not surprising considering that the training of scientific reasoning skills is mostly done within schools or other educational institutions. Others looked at the debate from a philosophy of science perspective. Some arguments are based on theoretical considerations, some originate from intervention studies.

Let us now begin by examining the arguments in favor of the domain generality of scientific reasoning. Inhelder and Piaget (1958) are amongst the most well-known proponents of domain generality. The scientific reasoning skills in the stage of formal operational reasoning were seen as independent from any content. Domain-general elements can be found in more recent conceptualizations, too. The scientific and engineering practices in the next generation science standards are supposed to be common across the sciences (NRC, 2012). Additionally, in the competence model and standards by Wellnitz et al. (2012) the four scientific reasoning skills *formulating questions*, *hypothesis generation*, *evidence generation*, and *evidence evaluation* are considered valid for physics, biology, and chemistry. Osbeck (2014) adds that all science consists of sense-making and that the three aspects of sense-making, namely inductive reasoning, explanations, and model-based reasoning, as well as processes like the selection of relevant facts, are common across many domains of science.

In addition to these theory-related arguments there is also empirical evidence in favor of the domain generality of scientific reasoning. Kuhn, Schauble, and Garcia-Mila (1992) targeted the control-of-variables strategy of children with a microgenetic intervention approach. Using the control-of-variables strategy means to only vary one target variable at a time while holding all other variables constant when trying to find out about the influence of

the target variable on an outcome. For instance, one task in this study was to determine which variables influence the speed of a race car (using a computer program). A child, who wanted to find out whether the engine of the car has an influence on the speed and who used the control-of-variables strategy, would use two cars that vary in engine size but are the same regarding the other variables that could be manipulated (color and presence of a muffler, presence of a fin, and wheel size). Besides the car task, children worked on tasks about how the features of a boat influence its speed and which kind of ball results in the best serve in a fictitious game. The results of the study demonstrated that certain scientific reasoning competencies are identifiable independent of the specific context domain.

Bridewell, Sánchez, Langley, and Billman (2006) created a tool that helps scientists with constructing and improving models based on data. The tool turned out to be useful in three domains within biology. The first domain was about predator-prey relations, the second about the influence of nutrients, light, and grazing behavior by zooplankton on the phytoplankton population, and the third about how photosynthesis regulation is influenced by various other variables (e.g. the amount of light or the concentration of mRNA). Halpern et al. (2012) included content from psychology, biology, and chemistry into their computer-based learning game for scientific reasoning. In the game students have to foil a plot by aliens, who want to confuse humanity by publishing flawed research. In order to achieve this goal students have to learn about 17 aspects of producing reliable research, e.g. the use of control groups, the influence of sample size, and conflicts of interest. The authors reason that the three included domains share common principles and that students can and should focus on these principles. Additionally, some intervention studies just assume the existence of domain-general skills without questioning that assumption (Dejonckheere, Van de Keere, Tallir, & Vervaet, 2013; Piekny & Maehler, 2013).

Next up are the arguments in favor of the domain specificity of scientific reasoning. H. H. Bauer (1994) connected his critique of a singular scientific method with the idea that the different sciences do not share relevant aspects. Combining philosophical and sociological views on science, Solomon (2001) followed a similar argument when he reported that the trend goes towards accepting that there is not a general scientific method. Kind and Osborne (2016) are also skeptical about any general scientific reasoning skills and view the sciences as distinctive, diverse, and not bound by a universal set of practices. Sinatra and Chinn (2012) see content knowledge as a requirement for reasoning. Tricot and Sweller (2014) even go as far as to say that all relevant knowledge is domain-specific. They do not deny the existence of domain-general skills but they claim that they are unteachable and thus of less interest compared to domain-specific skills. Additionally, in their view the application of a domain-general skill in a new domain is in itself a domain-specific skill. Yet another argument is that the validity of inferences in scientific reasoning is always bound to certain content and therefore these inferences are domain-specific (Brigandt, 2010).

Turning to empirical results, a study by Glaser, Schauble, Raghavan, and Zeitz (1992) showed that domain-general scientific reasoning strategies such as *evidence generation*, *data management*, and *evidence interpretation* were less important across tasks from physics and economics compared to abilities on a meta-level, e.g. selecting the right strategy. In the computer-based tasks students had to discover the underlying laws that determined the relationship between variables, e.g. Ohm's law about the relationship of current and voltage within a conductor or the law of demand that influences the relationship between the price and demanded quantity of goods. Furthermore, Schauble (1996) found in a microgenetic training study with children and adults that concept learning was very different in the knowledge-rich situations she used in this study compared to knowledge-lean situations. In knowledge-rich situations, theories about underlying causes for the relationship between

variables play an important role. In one of the two tasks that were used in the study, participants had to discover the ways in which four variables (boat size, boat shape, boat load, and canal depth) influence the speed of a boat being dragged through a canal. In the structurally-similar second task, in which objects were suspended from a spring, the dependent variable was spring extension, and the independent variables were object mass, object size, and how far the object was immersed in water. The observations Schauble made in these knowledge-rich situations led her to the conclusion that the contribution of domain-specific prior knowledge to scientific reasoning might be higher than other experimentation studies have suggested.

Apart from the more extreme positions there are also voices that see scientific reasoning as a combination of domain-general and domain-specific elements. This approach assumes that domain-general skills exist but that they are valid across only a few domains or that they are responsible for some but not all parts of the performance. The first of these more compromising conceptualizations became prominent about 30 years ago. K. W. Fischer and Farrar (1987) thought that skills in general are neither domain-specific nor domain-general but instead are learned in a domain-specific context and then gradually extended. It is possible for a person to have different skill levels in different contexts and with more advanced skills, generalizations become more likely. Similarly, Perkins and Salomon (1989) suggest, as a synthesis of voices in favor of general strategies and voices in favor of specific knowledge, that general skills exist and can be useful but that they work in contextualized ways. These arguments, which refer to skills in general, were mirrored by Klahr and Dunbar (1988) when it comes to scientific reasoning skills in particular: Their model contains both domain-specific and domain-general heuristics. For instance, knowledge is considered important, especially for *hypothesis generation*, but it is possible to set knowledge to the same level for everyone working on a task and observe domain-general skills at work. This

compromising stance was taken up by other authors and repeated in the following decades and continues to exist today (Erduran, 2007; Karmiloff-Smith, 2012; Niaz, 1995; Zimmerman, 2000).

Besides, it is possible to find supporting arguments for the compromising point of view in observational and intervention studies. In one training study the control-of-variables strategy had to be learned in a domain-specific context (Chen & Klahr, 1999). In the learning tasks third and fourth graders had to find out which variables influence how far a spring stretches, how far a ball rolls down a ramp, and how fast an object sinks in water. Once the learning task was mastered, fourth graders could transfer the skill to tasks with a biology or everyday life context, e.g. plant growth or baking cookies, respectively, and the skill helped them gain domain-specific knowledge in these other contexts. Schunn and Anderson (1999) gave a task from cognitive psychology to researchers from cognitive psychology (who had no expertise about the topic of the task), researchers from social and developmental psychology, and undergraduates. The participants had to determine which of two explanations for the spacing effect is correct. Both groups of researchers were better than undergraduates in the domain-general skills of designing an experiment and interpreting results, but there were also differences between the two groups of researchers in domain-specific skills like setting variables to realistic values. Thus, a successful performance depended on both domain-general and domain-specific skills.

Furthermore, a domain-specific science training for pre-kindergarten children, which aimed at the skills of observation and prediction, showed that these skills should be learned in a domain-specific context first, but after that they can be used in different contexts, even though this generalization is very hard (Gelman & Brenneman, 2004). An exemplary task from the training was to make observations about an apple and predict how many seeds it has. Another study made a historical analysis of how biologists have solved non-trivial

problems in their field (Scholl & Nickelsen, 2015). The authors found that the strategies used for discovery are to some extent content-neutral even though they are applied to domain-specific knowledge, which means that the discovery process as a whole is an interaction of domain-general strategies and domain-specific knowledge.

There are two more aspects of this debate that should be mentioned. First, even some of the arguments for domain generality or domain specificity are not as clear cut as they first seem. For instance, the Kuhn et al. (1992) study acknowledges that the domain-general skills emerged from a domain-specific context and Kind and Osborne (2016) state that some of their styles of reasoning are useful in various disciplines and that they have to be learned from examples, which implies that the learner can transfer them to other contexts. Second, the domain generality vs. domain specificity debate is not only fought within the area of scientific reasoning. In the research literature about NOS one can find studies in favor of domain generality (Abd-El-Khalick, 2012), studies in favor of domain specificity (Schizas, Psillos, & Stamou, 2016), and also the compromising point of view (Muis, Bendixen, & Haerle, 2006). This highlights the relevance of the problem and it might mean that contributing to the debate around the domain generality of scientific reasoning also benefits other areas of research.

The debate around the domain generality vs. domain specificity of scientific reasoning is not only relevant from a theoretical point of view. There are also direct consequences for the assessment of scientific reasoning. One contested aspect of the assessment of scientific reasoning is whether it is possible to reduce context effects in an amount that it becomes possible to assess scientific reasoning in a domain-general way. Harlen (1999) is in favor of the possibility of doing so. She argues that the assessment of scientific reasoning will be influenced by content but that does not mean that it will be dominated by content. It is possible to reduce the content-specific influence by using content available to students.

According to her, too much focus on domain specificity would even be detrimental to scientific reasoning skill assessment because tasks might appear harder than they actually are due to domain-specific aspects. Other researchers argue that knowledge-lean tasks are not useful, that they are just circumventing the problem that knowledge will play an important role in realistic scenarios, and that knowledge and reasoning should not be separated (Kind, 2013; Osborne, 2013; Zimmerman, 2000). From the perspective of test construction it is also very relevant whether scientific reasoning is seen as domain-general or domain-specific. If scientific reasoning is domain-general we can construct one test that will be applicable to many populations and content areas. However, if scientific reasoning is domain-specific different tests have to be developed for various populations and with different context domains.

In this thesis, Study 1 explores how the domain generality vs. domain specificity debate is settled in assessments and if there are approaches that include more compromising ideas. Study 2 picks up the idea that the debate can be phrased as a problem of measurement invariance and thus a problem of construct validity. The concepts of measurement invariance and construct validity will be explained in more detail in the next section in the part about important psychometric properties.

1.3 The Evaluation and the Evolution of Scientific Reasoning Tests

1.3.1 Important psychometric properties

When evaluating tests, it is crucial to take their psychometric properties into account. The next paragraphs give an overview of the most important psychometric properties and how they play a role in evaluating scientific reasoning tests in general and in Studies 1 and 2

in particular. In addition, the next paragraphs will cover the debate surrounding the concept of validity.

The main psychometric properties are objectivity, reliability, and validity, because if they are lacking, a meaningful interpretation of test results is not possible. For objectivity and reliability textbooks agree on their definition (Moosbrugger & Kelava, 2008; Rost, 2004). Objectivity, often separated into objectivity of the administration, objectivity of the scoring process, and objectivity of the interpretation, deals with the question of whether test results are independent from influences apart from the test taker, e.g. the people who administer, score, and interpret the test. With the concept of interrater agreement and the indicator Cohen's kappa, standard ways exist to evaluate this psychometric property. Objectivity will not play a major role in this thesis. It should be noted, though, that scientific reasoning tests that require the coding of open-ended questions should always check their objectivity.

Reliability addresses the question of which part of the total variance can be attributed to differences in the latent variable. There is an agreement about this reliability definition even if different measures are used to evaluate reliability (Revelle & Zinbarg, 2009). Reliability will get some attention in both studies.

In comparison to reliability there is a bigger debate around the concept of validity, which is why we spend more time making clear what is meant by validity in this thesis. Validity asks the question of whether the test is measuring what it is supposed to measure. The main focus in this thesis will be on this psychometric property, because validity is considered to be the most important psychometric property (Moosbrugger & Kelava, 2008; Rost, 2004): Only if a test is valid it is possible to infer the level of the latent variable that was intended to be measured from the performance in a test. Thus only a valid test makes it possible, via a latent variable, to generalize from the performance in the test situation to the performance outside the test situation, which is often the ultimate goal of testing. In this

thesis, we will distinguish the following subparts of validity: content validity, construct validity (including factorial structure, divergent validity, and convergent validity), and criterion validity. This is a common distinction of validity aspects (see e.g. Moosbrugger & Kelava, 2008). Discerning different parts of validity will be useful for this thesis. It allows us to make more specific diagnoses of single tests and the field of scientific reasoning assessment. Instead of just saying that a test or the field of scientific reasoning assessment as a whole lacks validity, it becomes possible to say which specific validity aspects are lacking, which will also lead to more specific recommendations for the future.

Content validity asks if the test items are representative of the construct. Usually content validity is not expressed with a number and instead is based on an evaluation by experts. It is important, but of course also hard, to compare between tests.

Construct validity addresses the question of whether a test has a sound theoretical grounding: Is the variation in the intended latent variable really the source for the variation in the test? Is it thus adequate to draw inferences from the test score about a latent variable? One way to evaluate construct validity is via factor analysis. In addition to the factorial structure, construct validity is also determined by convergent and divergent validity. These two latter validity aspects deal with the placement of the latent construct into a nomological net of other constructs. A nomological net shows the relation of a target construct to other constructs. It states which other constructs overlap with the target construct and which are considered to be closely related but distinct constructs (Ziegler & Hagemann, 2015). It is possible to infer expected correlations between different tests from the supposed position of the target construct within the net. Convergent validity is understood as the correlation of a test with other already established tests for the same construct. Divergent validity is supported by low or non-existent correlations with tests that claim to measure constructs that are close to the measured one but are not supposed to be the same. If both convergent and divergent validity

are high, clear borders can be drawn between those other constructs and a well-defined target construct. In the case of scientific reasoning, there seem to be two obvious constructs that should be considered in establishing divergent validity: general reasoning and science knowledge.

In addition to factorial structure, convergent validity, and divergent validity, some scholars see measurement invariance as a part of construct validity (Borsboom, Mellenbergh, & Van Heerden, 2002). Others categorize the question of measurement invariance under the aspect of fairness (Moosbrugger & Kelava, 2008). In this thesis, the former approach will be adopted, because a test that is measurement invariant possesses the same factorial structure in two populations. However, if measurement invariance is violated and group membership (e.g. determined by gender or race) leads to biased test results, it is not certain anymore that the latent variable that was intended to be measured is the sole source for variation in the test, which is exactly what construct validity is about. Probably more important than the categorization of measurement invariance, though, is that it gets checked at all, which is done in Study 2.

The last validity aspect is criterion validity, i.e. the question of whether a test can predict a variable outside of the test situation. For instance, a scientific reasoning test could be validated by predicting academic performance in science classes. If the criterion that is measured is a future performance, criterion validity is also called predictive validity.

These notions about validity are not uncontested. Debates about the concept of validity often focus on construct, criterion, and predictive validity. Borsboom, Mellenbergh, and van Heerden (2004) see construct validity as the only important aspect of validity because it deals with the relation between test scores and a latent variable. Thus, they argue that construct validity addresses a question of causation, namely whether the variance in the test scores is caused by the variation in a latent variable. In contrast, criterion validity is just about

correlation, depends on the specific criterion you choose, and can vary when you exchange the independent with the dependent variable (Molenaar & Borsboom, 2013; Rost, 2004). Phrased differently, criterion validity is more concerned about the practical value of a test than the accuracy of the test's claims about an underlying theoretical construct. Consequently, Borsboom & Mellenbergh (2007) differentiate between measurement concepts (construct validity and reliability) and decision concepts (predictive accuracy). Similarly, Millsap (2007) differentiates between measurement invariance, which he defines as the absence of bias for specific groups in the test itself, and predictive invariance, which is about equal predictive accuracy of a test for a criterion between groups. Furthermore, these authors argue that measurement invariance should be preferred to predictive invariance (Molenaar & Borsboom, 2013), that measurement invariance is not getting enough attention (Millsap, 2007), and that the two are even mutually exclusive under realistic conditions (Borsboom, Romeijn, & Wicherts, 2008). Additionally, construct validity often benefits from a homogeneous test while criterion validity often benefits from a heterogeneous test (Rost, 2004).

However, there are also voices that emphasize the importance of predictive performance for validity and that reliability and construct validity are not the only relevant aspects in a test evaluation (Blömeke, Gustafsson, & Shavelson, 2015). This thesis will follow the latter approach. Thus, Study 1 gives an overview of the extent to which different psychometric properties, mainly a range of validity aspects, are checked in current scientific reasoning tests. While construct validity is certainly of very high importance and receives a lot of attention in Study 2, a full evaluation of a test should also consider criterion validity. That is why criterion validity will play a role in both studies.

After discussing psychometric properties on a rather abstract level, let us take a look at how some of these topics are relevant to the assessment of scientific reasoning. An important

aspect in terms of construct validity is the structure of scientific reasoning, mainly its dimensionality. Are the different skills parts of one general scientific reasoning ability or are they independent of each other? Can scientific reasoning be assessed with a single measure that might leave out some aspects? This would only be feasible if the single skills are just some possible realizations of a universe of scientific reasoning facets, which would make the single skills interchangeable indicators of scientific reasoning. However, if the skills are independent we have to measure all of them or risk ending up with an incomplete picture of scientific reasoning.

In terms of divergent validity, the distinction from general reasoning is especially important when arguing for the domain generality of scientific reasoning, because general reasoning is also conceptualized as a domain-general construct. However, the connection between scientific reasoning and general reasoning is not clear yet. There are people who think that scientific reasoning has nothing interesting to offer as a concept beyond general reasoning (Simon, 1966). In contrast, demonstrating the difference between scientific reasoning and science knowledge is especially important when arguing for domain specificity. Embedding a construct into a nomological net is also very important for future test construction in general, because it makes it easier to avoid unwanted multidimensionality (Ziegler & Hagemann, 2015). Of course this can be hard if the net is woven very tightly, which might just be the case with scientific reasoning since there are so many closely related constructs, such as general reasoning, scientific argumentation, NOS, and science knowledge.

While some might see teaching scientific reasoning as having merit on its own (Harlen, 1999), it is widely regarded as important to explore the criterion validity of scientific reasoning. The predictive validity of tests is of special interest. It would speak for the quality and relevance of a test if, for instance, its results can predict how well students can acquire content knowledge or how well they will do in their academic careers.

1.3.2 The Rasch model and differential item functioning

Since item response theory in the form of the Rasch model and differential item functioning (DIF) play a central role in Study 2, the two terms will be briefly introduced in the following paragraphs. The rationale of the Rasch model was introduced by Georg Rasch more than half a century ago (Rasch, 1960). The great advantage of the model is that person and item parameters get separated (Fisher, 2004). It is this property that potentially, i.e. as long as the data fit the model, allows judging the ability of a person, independent of the specific items (from a larger set of items that all measure the same construct) that were used, as well as judging the difficulty of the items, independent of the specific people who took a test (Rost, 2004). Another advantage is that the Rasch model (and also other models based on item response theory) allows mapping dichotomous answer data onto a continuous latent variable (Fisher, 2004). A further advantage compared to classical test theory is that the fit of the data to the model can be assessed. When there is a lack of fit, there are still lessons to be learned from that situation, e.g. about which items might be the source of the misfit and if the construct needs to be revised or even abandoned because it was falsified (Heene, 2007). Rasch modelling is relevant for the assessment of scientific reasoning because if we would consider basing a high-stakes decision on a scientific reasoning test, e.g. the selection of students for a graduate program, the test should conform to a Rasch model (Heene, 2007).

A special feature of the Rasch model is that it allows tests for DIF, which indicates the absence of measurement invariance. The basic idea of a test for DIF is that two persons with the same level on the latent construct should perform similar on test items (a difference between these two persons can still occur due to measurement error). Thus, in order to test DIF, members of two different groups are matched according to their test score and it is then calculated if differences in the difficulties of items appear for the two groups, which would indicate DIF (Zumbo, 1999). Boone, Staver, and Yale (2014) agree with the aforementioned

notion that measurement invariance, with DIF as an indicator, touches on construct validity, because if DIF is present it implies that more than one construct is measured by a test. In this thesis, Study 2 demonstrates how the idea of DIF can be useful to the debate of the domain generality vs. domain specificity of scientific reasoning.

1.3.3 A brief history of the assessment of scientific reasoning skills

This subsection traces the history of scientific reasoning assessment. The subsection functions to demonstrate two aspects: First, it will show how assessments have changed over time. Second, using concrete examples instead of theoretical notions, it will show how the debated aspects of scientific reasoning, which were mentioned all through this Introduction, are reflected in scientific reasoning tests. The debated aspect that will be most obvious is the question of which skills are part of scientific reasoning. There are also hints at other current problems such as the controversy around domain generality, the connection with general reasoning, and whether a test is an accurate reflection of scientific reasoning. A focus of this history will be on the beginnings of scientific reasoning assessment since more recent tests are covered in more detail in Study 1.

The first noteworthy test was made by Herring (1918). The 11 so-called scientific thinking abilities that are covered in the test contain some tasks that are vaguely reminiscent of modern scientific reasoning tests. For instance, test takers have to select facts that are relevant for a question, which can be seen as a form of gathering evidence. However, test takers also have to distinguish between definitions with varying quality and to correct an unclear sentence, tasks that would be hard to categorize into one of the eight scientific reasoning skills in Table 1. It is striking that the test is heavily language dependent and some answers seem subjective, e.g. when test takers have to decide which question could be solved best by using statistics. As opposed to recent tests, this test places itself into the same

category as tests for intelligence and memory. The author notes that it is an unresolved question whether the tested abilities can be transferred from geography, the context domain of the test, to other subjects, an issue which relates to the question of domain generality. Additionally, the test author admits that it is unclear and that it should be determined in the future whether the test actually measures what it intends to measure.

More tests with items that resemble the current understanding of scientific reasoning skills were created around 1940. The Stanford Scientific Aptitude Test for High School and College Students (Zyve, 1939) contains items that target the control-of-variables strategy, which is also a common subject of current scientific reasoning tests and interventions. However, there are also items dealing with estimation or optic illusions. This test specifically states that it is not about intelligence but instead about finding the best fit of a person to a career. A test by Blair (1940) also seeks to measure scientific thinking, and while it contains tasks that are reminiscent of *drawing conclusions*, there are also subparts of the test about intellectual honesty or open-mindedness. There are definitely some items that seem strange from a modern perspective, e.g. questions about whether a high forehead is a sign of intelligence or whether any nation that persecutes Jews is uncivilized. An evaluation of the test showed that scientists do not agree with many answers that are correct according to the scoring rubric (Blair, 1940). Yet another test that claims to measure scientific thinking was created by Engelhart and Lewis (1941). Interestingly, the authors state that the test might not deserve this label, although it probably is the test that comes closest to our current understanding of the term, with test takers rating how a certain piece of evidence is related to a hypothesis and drawing conclusions.

A last test from the early period that should be mentioned is by Macy and Wood (1951). Again, we can find elements of scientific reasoning with a task about making evidence-based judgments. However, the 113 items also cover interests and superstitions. In

one question test takers have to judge whether scientists might one day tell the sex of a child before birth and in another question it is considered true that fortune tellers may have the ability to tell something about the future. A common theme in these older tests is that they mix some items measuring something we might classify as a scientific reasoning skill today with a wide range of other aspects. Additionally, most of these tests relied heavily on verbal skills.

When we consider these old tests it becomes more obvious how big of a leap the theory by Inhelder and Piaget (1958) was. Several tests that were developed in the decades after the publication of their work were built on their theory and these tests have a much stronger resemblance to current scientific reasoning tests. The most prominent example was the Classroom Test of Scientific Reasoning (CTSR; Lawson, 1978), which is still used today. This test is explored in Study 2. Using the CTSR as an example, the study also shows that the problem of which skills are part of scientific reasoning had not been settled despite the obvious advances that were made.

Especially in the last two decades, standards by educational institutions and results from large-scale assessments like PISA (OECD, 2006) have played a very prominent role in shaping the assessment of scientific reasoning. The national science education standards (NRC, 1996) were a prime example of an educational institution valuing scientific reasoning skills. The standards emphasized scientific literacy and scientific inquiry skills. For grades 9-12 they mentioned the following skills: *identify questions*, *design and conduct investigations*, *use mathematics and technology*, *formulate and revise scientific explanations*, *recognize and analyze alternative explanations*, and *communicate and defend a scientific argument*. However, the debate about which skills belong in a test about scientific reasoning skills has not abated. Even today, tests that claim to measure scientific inquiry still actually measure content knowledge with a majority of their questions (Day & Matthews, 2008).

It is clear, though, that the assessment of scientific reasoning and the problems surrounding it get more attention these days. The 21st century challenge that was mentioned at the beginning of this thesis is to gather more knowledge about the scientific reasoning construct and expertise about how to assess scientific reasoning (Osborne, 2013). In addition, the NRC accompanied their newest standards with a report titled "Developing Assessments for the Next Generation Science Standards" (Pellegrino, Wilson, Koenig, & Beatty, 2014). The report shows that the organization recognizes the importance, but also that it is concerned about the current state, of scientific reasoning assessment.

1.4 Research Goals and Questions

There are several reasons why it is important to find out more about all of the aspects of scientific reasoning assessment that were mentioned in the last sections. As was shown at the beginning of this thesis, scientific reasoning is considered an important skill set. It was also shown that scientific reasoning can be trained (Engelmann et al., 2016). However, a necessary requirement for establishing and comparing interventions is a precise measurement device. Additionally, the creators of tests have several responsibilities. The design of a test itself can influence concept knowledge (C.-Y. Chang, Yeh, & Barufaldi, 2010). Pellegrino (2013) and Shavelson and Huang (2003) agree that assessments have an influence on what is taught, so if we are not careful our tests might set the wrong incentives for science education. For instance, the way that the domain generality vs. domain specificity debate is settled in the assessment of scientific reasoning will influence whether the focus of teaching is on domain-specific content, domain-general skills, or a combination thereof.

With the importance of scientific reasoning assessment and the debated aspects of scientific reasoning in mind, this thesis set out to learn more about this complex field. The

first main goal in this task was to explore the current state of scientific reasoning assessment and scientific reasoning conceptualizations. The following questions were aligned with this goal:

1. Which assessments of scientific reasoning currently exist?
2. How do they conceptualize scientific reasoning?
3. Particularly, which skills are part of scientific reasoning, how are the connected, and are they considered to be domain-general or domain-specific?
4. What do the tests do to control psychometric properties?
5. Did any of these aspects change over time?

Answering the first question will mostly yield practical benefits to researchers and practitioners in need of a test. More importantly, the answers to the other questions will contribute to the debates around the construct of scientific reasoning, which were described in this Introduction. The way these debates are settled, especially regarding the conceptualization of scientific reasoning, will in turn have consequences, for instance on how we design curricula and trainings for scientific reasoning.

The second main goal of the thesis was to show how an in-depth analysis of an already existing scientific reasoning test can help us learn about the scientific reasoning construct and what has to be improved in the construction of future scientific reasoning tests. The questions that went along with this goal were:

1. Is it possible to assess scientific reasoning in a domain-general way?
2. What is the construct validity of scientific reasoning, considering both factorial structure and divergent validity?
3. What is the criterion validity of the test?
4. Are the deployed statistical methods useful in evaluating scientific reasoning tests?

The answers to these questions are again important from a practical, theoretical and methodological perspective. Practitioners can find out how much trust they can place on the results from this specific scientific reasoning test and, to a certain extent, on results from scientific reasoning tests in general. Much like the results from Study 1, Study 2 will contribute to the debates around the conceptualization of scientific reasoning. Additionally, it will produce recommendations about how we should evaluate and construct scientific reasoning tests.

1.5 Overview of Studies

Study 1 is dedicated to the goal of exploring the current state of scientific reasoning assessments and conceptualizations. To contribute to that goal a review of scientific reasoning tests was conducted. The tests that were found were analyzed regarding which scientific reasoning skills they included, which theories were used for test construction, the stance the authors took on the domain generality vs. domain specificity debate, the controls of psychometric properties, and the test format that was used, as well as other minor aspects of the tests. The study also looked at trends over time.

Study 2 is dedicated to the goal of learning from an in-depth evaluation of an existing scientific reasoning test. The study focusses on the scientific reasoning of university students. Using one of the most used scientific reasoning tests (Lawson, 1978), Study 2 deployed a combination of older and newer test analysis methods to explore the domain generality of the whole test and of single items, the connection with general reasoning and content knowledge, the factorial structure of the test, and the criterion validity of the test.

2 Study 1 – A Review of Scientific Reasoning Tests

2.1 Introduction

A main goal of science education in national and international guidelines is to enable students to use scientific concepts and methods to address problems in research, professional practice, and daily life (Abd-El-Khalick et al., 2004; NRC, 2012; OECD, 2006). These scientific concepts and methods are seen as necessary for inquiry learning (Lazonder & Harmsen, 2016), as one part of science education that is needed for civic engagement (Rudolph & Horibe, 2016), and as a vital part in preparing a competitive workforce (The Royal Society, 2014). Typical examples for these concepts and methods are the skills[2] to construct an experiment, to test a hypothesis, or to draw conclusions from tabulated data.

These and similar skills can be found in different concepts. Almost all conceptualizations of scientific literacy acknowledge that scientific literacy consists not only of knowledge but also of skills (Norris et al., 2014). Pedaste et al.'s (2015) scientific inquiry model includes sub-phases such as *questioning*, *hypothesis generation*, *experimentation*, *data interpretation*, and *communication*. The OECD framework for PISA (OECD, 2006) includes three skills: *identifying scientific issues*, *explaining phenomena scientifically*, and *using scientific evidence*. In this review we focus on the eight skills shown in Table 2.

[2] We understand a skill in a general sense as being distinct from both intelligence (because a skill can be trained) and from conceptual knowledge. We use the rather general term "skill" to incorporate the different terms used in different scientific reasoning conceptualizations.

Table 2

List of Scientific Reasoning Skills Used in This Review

Skill name	**Skill description**
Problem identification	Perceiving a mismatch between a problem (from a science, professional, or real-world context) and current explanations, analyzing the situation, and building a problem representation.
Questioning	Identifying one or more questions as the basis for an upcoming reasoning process.
Hypothesis generation	Constructing possible answers to a question (according to scientific standards) based on known models, frameworks, or evidence.
Construction and redesign of artefacts	Creating a prototypical artefact (e.g. an engineer building a machine or a teacher constructing a learning environment), testing it, and revising it based on the test.
Evidence generation	Producing evidence following one of several methods. Amongst them are controlled experiments, observational studies, and deductive reasoning based on a theory.
Evidence evaluation	Analyzing various forms of evidence in regard to a claim or theory.
Drawing conclusions	Coming to a conclusion by weighing the relevance of different pieces of evidence. Can lead to the revision of an initial claim.
Communicating and scrutinizing	Presenting and discussing the methods and the results of a scientific reasoning process both within a team and a broader community.

Note. Adapted from F. Fischer et al. (2014).

We summarize these skills under the term *scientific reasoning*. For the purpose of this review we regard scientific reasoning as being different from the *nature of science* construct (Lederman et al., 2002) as it is focusing on knowledge about science. This review will also exclude (collaborative) scientific argumentation, which is a complex process of its own (Berland & McNeill, 2010). Scientific reasoning focuses on individuals rather than on their interaction. While discussions are part of many scientific reasoning conceptualizations, they are placed at the end of a process or accompany other core steps, but they are not at the core of the model itself (Pedaste et al., 2015).

The origins of the concept of scientific reasoning skills go back several decades and older conceptualizations exist, which still influence how we think about scientific reasoning. The most advanced stage of Inhelder and Piaget's (1958) theory about the stages of the development of human thinking, formal operational reasoning, includes an important aspect of scientific reasoning: Children on this level are supposedly able to use evidence to evaluate hypotheses. Klahr and Dunbar (1988) developed another prominent conceptualization in their Scientific Discovery as Dual Search (SDDS) model which contains in its cyclical structure the three research phases *hypotheses generation*, *evidence generation*, and *evidence evaluation*. While Piaget was assuming a single cognitive ability that is generally applicable, the conceptualization by Klahr and Dunbar moves away from this idea. The research phases are part of a problem-solving process but they are distinguishable and therefore possess a certain degree of independence. Additionally, while not abandoning the idea of domain-general aspects, the SDDS model is also emphasizing the important role of domain-specific prior knowledge in the scientific reasoning process. One reason for this shift from domain generality towards domain specificity probably was the growing focus on the interaction of general skills with domain-specific knowledge. Perkins and Salomon (1989) name several examples for this interaction. For instance, it is a generally useful strategy to think of counterfactuals to evaluate a claim. However, domain-specific knowledge is needed to construct valid counterfactuals in a specific domain.

The differentiation into a higher number of independent skills continues in newer conceptualizations of scientific reasoning, where it is common to consider at least eight skills as contributing to scientific reasoning, such as defining problems, formulating questions and hypotheses, gathering and evaluating evidence, and explaining and communicating results (F. Fischer et al., 2014; NRC, 2012). To summarize, the differences between conceptualizations of scientific reasoning that exist are in (a) the skills they include, (b) if there is a general,

uniform scientific reasoning ability or rather more differentiated dimensions of scientific reasoning, and (c) if they assume scientific reasoning to be domain-general or domain-specific.

The inclusion in recent educational guidelines and large-scale assessments shows that there is a continued interest not only in the construct of scientific reasoning itself but also in its measurement. The aspect of measurement is an important one considering that only well-constructed measurement instruments with well-known psychometric properties can be the basis for effective interventions and informed policy decisions (Wiliam, 2010). However, so far we lack information about what scientific reasoning tests exist and how they conceptualize and assess scientific reasoning. Additionally, we should know to what extent the tests explore the similarity with other test instruments and if their results can predict other variables of interest like academic success. We therefore conducted a review of scientific reasoning tests with two goals in mind. First, the review should give an overview of existing measurement instruments that claim to measure scientific reasoning. Second, the review analyses issues of both theory related relevance (how is scientific reasoning conceptualized) and practical relevance (e.g. information about test formats and target groups). To address the issues of theory related relevance we made the assumption that the conceptualizations made by test authors can be used as a proxy for general developments of differences in scientific reasoning conceptualizations. By striving for these two goals we hope that the review is relevant to different groups: For researchers and practitioners looking for a scientific reasoning test that fits their purpose but also for researchers who are interested in how tests of scientific reasoning reflect the differences in conceptualizations of scientific reasoning we described above. Besides, researchers thinking about creating a new scientific reasoning test can learn from the shortcomings of current tests that we present in this review.

These were the questions we wanted to answer: First, which scientific reasoning skills are addressed by the tests that intend to measure scientific reasoning? As we have shown above there are different conceptualizations emphasizing different skills and it is unknown if this is reflected in tests. Are some skills considered as more important and is there a lack of tests for others? Second, which theoretical frameworks are used for test construction? Are different scientific reasoning skills connected to each other via an underlying ability or are they independent, according to these theoretical frameworks? To put it another way, is scientific reasoning conceptualized as a single (i.e., unidimensional) competence with facets or rather as a set of relatively independent skills (multidimensional)? The answer to this question has important consequences for both measuring and fostering scientific reasoning; namely, if it is possible to test and facilitate a single skill that will sufficiently represent the absent aspects or whether different skills need independent measurement and instruction. Third, we address the question of how closely tied to a specific domain the tests conceptualize scientific reasoning. So, is scientific reasoning regarded as domain-specific or rather domain-general (independent of the question of how many dimensions there are)? Do we have to draw a distinction, for instance, between scientific reasoning in chemistry, biology, physics, and non-science fields – and, if so, in which way does scientific reasoning differ? Knowing this might in turn influence the scope of future tests and inform us if teaching scientific reasoning in one domain helps a student with scientific reasoning in another domain. Fourth, what can be said about the psychometric properties of current tests? If we want to base high-stake decisions on test results, for instance, the access of students to a graduate program, we also need to focus more on this aspect of tests. We would certainly like to know the relation to other scientific reasoning measures and different but related concepts like general cognitive abilities and the relevance of results outside of the test context. Tests of psychometric properties, e.g. tests of construct validity, might also contribute to the

discussions about different conceptualizations of scientific reasoning, especially the question of dimensionality. Thus, it is important to find out if these issues are currently addressed by test authors. *Fifth*, how do the test instruments approach the measurement of scientific reasoning? This is probably especially relevant for researchers and practitioners who are looking for a scientific reasoning test. One researcher might need a short measurement that can be used as one amongst many measurements in a study whereas a practitioner might look for a test with higher ecological validity. For people looking for a test it is important to know what options they have. Last, with all of these questions we analyzed if we can observe any trends over time.

2.2 Methods

2.2.1 Literature search

We conducted a literature search using the databases ERIC, PsycINFO and PSYNDEX. Search strings were all possible combinations of the terms "scientific reasoning", "scientific thinking", "scientific literacy", "scientific inquiry", "scientific discovery" or "science process skill*" together with the terms "test", "assess*", "measur*" or "scale". We used this variety of search terms because, as we have explained in the Introduction, the skills we are interested in are included in concepts that can go by different names. In addition to this search, references in tests selected for the review were considered. The search took place between October 2013 and June 2014. After sighting the results returned from these search procedures we subjected 84 sources to a closer analysis in order to decide on their inclusion into the review.

The following inclusion criteria were then used to select tests: First, at least one of the descriptions by the test authors of what the test is measuring could be related to one of the

scientific reasoning skills of an interdisciplinary conceptualization by F. Fischer et al. (2014). Its eight skills are mentioned in Table 2. It was selected because of its inclusive nature: It was created by 12 professors from various disciplines (psychology, education, biology, medicine, mathematics, media informatics and social work) so it should also be applicable to the conceptualizations used by tests from different disciplines. Besides, it overlaps completely or almost completely with many other scientific reasoning conceptualizations. For instance, the three research phases of the SDDS model (Klahr & Dunbar, 1988), *hypotheses generation*, *evidence generation*, and *evidence evaluation,* are also part of the conceptualization by F. Fischer et al. (2014). Another example is that both the conceptualizations by F. Fischer et al. (2014) and the NRC (2012) include the skills of defining problems, formulating questions and hypotheses, gathering and evaluating evidence, and explaining and communicating results. Because of these overlaps with many other conceptualizations and its interdisciplinary nature, the conceptualization by F. Fischer et al. (2014) seemed like a good basis for a review.

Second, the instrument had to be a test that could be and was intended to be used beyond a single study. A necessary indicator for this criterion were reports on either content validity, construct validity, criterion validity, or norms. No constraints were made regarding the publication date. After applying the inclusion criteria, 46 sources were excluded. The main reasons to exclude a source were that the whole test measured something else (e.g. knowledge), not enough information could be gained from the text to be sure that the inclusion criteria were met, the source was not about a test (but rather e.g. about an intervention) or no reports on validity or norms were given.

2.2.2 Test analysis

The selected tests were analyzed regarding their year of development, target group(s), addressed skills, theoretical background(s) including the dimensionality of their structure, domain generality vs. specificity assumptions, certain psychometric properties (reliability, content validity, construct validity – also including convergent and divergent validity – , criterion validity, and norms), and test format. The year of development refers to the year in which the test was first used if the authors provided this information. If not, the year of development is identical to the year of the (first) publication about the test. For large-scale assessments the year of the introduction of the most recent framework for which the data analysis is already completed was used. For instance, PISA introduced a new science framework in 2006 that is the most elaborated science framework until now (with a completed data analysis). Thus, PISA was assigned with 2006 as the year of development for this review.

To determine which skills were measured by which test, we started by extracting short descriptions of the tested skills from the original articles. Then, the first author and a second rater coded the descriptions based on a coding scheme that was developed on the basis of the scientific reasoning conceptualization by F. Fischer et al. (2014)[3]. Consequently, descriptions were sorted as representing one of the eight skills or into an "other" category if they did not fit into the eight skill categories. Descriptions consisted of one or a few words, in some cases of a complete sentence. For instance, the description "formulating and judging ideas / hypotheses" was coded as *hypothesis generation* and the description "data analysis" was coded as *Evidence evaluation*. Overall, 258 descriptions of skills were sorted. Roughly 10% of the data were used as examples for the coding scheme. Two training rounds used roughly

[3] The coding scheme is included as Appendix A in this thesis.

15% of the data each. To determine inter-rater reliability Cohen's Kappa was calculated with the remaining 60% of the data. An agreement of .792 was achieved. In cases of disagreement, agreement was reached by a discussion of the two raters.

Theories used for test construction were analyzed by all three authors of this review. The first author of this review wrote summaries of the theories based on the descriptions of the theories that the test authors referred to in their articles. Based on these summaries all three authors of this review discussed the theories in respect of their dimensionality. Three categories became apparent through these discussions: unidimensional theories – assuming one general ability developing over time (e.g. Inhelder & Piaget, 1958) –, multidimensional theories – postulating several independent skills (e.g. Livermore, 1964) –, and theories assuming a problem-solving process consisting of multiple skills (e.g. Klahr & Dunbar, 1988). Therefore, theories were sorted into one of these three categories. The determination of the domain generality vs. specificity aspect was based on claims by the test authors they made in their publications. If statements from the authors made it clear that they assume either an overarching domain-general construct or that the skill set measured by the test is inextricably connected with a domain, these tests were categorized as assuming domain generality or domain specificity, respectively. If the authors made no such assumptions or their assumption could not be clearly evaluated, the according tests were sorted into a third category. Additionally, there was a fourth and last option for the rating of the domain generality vs. specificity aspect: A test was put into this category if a test author made the assumption that certain parts or subscales of the test are general but others are domain-specific. In most cases it could be easily determined from the descriptions by the authors if they were checking reliability, content validity, construct validity – also including convergent and divergent validity –, or criterion validity. In case it was not clear which psychometric property was checked by the test authors, the uncertainty was resolved through a discussion

of the first and second authors of this review. For validity checks, it was also noted which measures were used to establish validity.

2.3 Results

2.3.1 Overview

In total, we found $k = 38$ tests that fulfilled the inclusion criteria. For a complete list of these tests and an overview of their characteristics see Table 3. The tests were developed in two waves: 11 tests were developed between 1973 and 1989 and 27 tests were developed between 2002 and 2013. The main target populations were secondary school students ($k = 22$), followed by college and university students ($k = 14$), and elementary school students ($k = 12$; tests can have more than one target group). Only four tests targeted populations other than the above and only two of these tests targeted populations outside educational institutions.

Table 3

Overview of Tests Included in the Review and a Selection of Their Properties

Test name	References	Test format[a]	Covered scientific reasoning skills[b]	Target group(s)[c]	Assumption about domain generality[d]	Context domain(s)[e]	Checks of psychometric properties[f]	Test norms[g]
A written test for procedural understanding	(Roberts & Gott, 2004)	**OP**	**EG, EE, DC, CS, OT**	**S**	**s**	**B, C, P**	**R, CS, CR**	**-**
Abilities in scientific inquiry	(Nowak, Nehring, Tiemann, & Upmeier zu Belzen, 2013)	**MC**	**Q, HG, EG, EE, DC, OT**	**S**	**s**	**B, C**	**R, CS**	**-**
Assessment of Critical Thinking Ability (ACTA) Survey	(White et al., 2011)	**MI**	**EG, DC**	**U**	**g**	**M**	**CR**	**-**
Assessment of Scientific Thinking in Basic Science	(Azarpira et al., 2012)	**MI**	**HG, EG, EE, DC, OT**	**U**	**n/u**	**M**	**CS, CR**	**-**
Chemistry Concept Reasoning Test	(Cloonan & Hutchinson, 2011)	**MC**	**EE, DC, OT**	**S, U**	**s**	**C**	**CT, CR**	**-**
Classroom Test of Scientific Reasoning (Lawson-test)	(Lawson, 1978)	**MC**	**EG. EE, OT**	**S, U**	**g**	**na**	**R, CT, CS, D/C, CR**	**+**
Competence Scale for Learning Science	(H.-P. Chang et al., 2011)	**SA**	**Q, HG, EG, EE, DC, CS**	**E, S**	**n/u**	**na**	**R, CT, CS**	**-**
Constructive Inquiry Science Reasoning Skills (CISRS)	(Weld, Stier, & McNew-Birren, 2011)	**OP**	**HG, EG, EE, OT**	**U**	**g**	**na**	**CT, D/C**	**-**

Test name	References	Test format[a]	Covered scientific reasoning skills[b]	Target group(s)[c]	Assumption about domain generality[d]	Context domain(s)[e]	Checks of psychometric properties[f]	Test norms[g]
Detector - Inquiry Intelligent Tutoring System	(Gobert et al., 2013)	**AA**	**HG, EG, EE, DC, CS**	**S**	**s**	**B, ES, P**	**CT, CS**	**-**
Empirical based reasoning	(Heene, 2007)	**OP**	**EG, EE**	**U**	**s**	**BS**	**R, CS, CR**	**-**
Evidence-Based Reasoning Assessment System (EBRAS)	(Brown, Nagashima, Fu, Timms, & Wilson, 2010)	**OP**	**DC, CS, OT**	**S**	**g/s**	**P**	**R, CS, CR**	**-**
Experimental Design Ability Test (EDAT)	(Sirum & Humburg, 2011)	**OP**	**EG, OT**	**U**	**g**	**na**	**CR**	**-**
Experimental problem-solving	(Ross & Maynes, 1983)	**MC**	**HG, EG, EE, DC**	**S**	**g**	**na**	**R, CT, CS, D/C, CR**	**-**
Experimenting as problem-solving	(Hammann, Phan, & Bayrhuber, 2008; Hammann, Phan, Ehmer, et al., 2008)	**MC**	**HG, EG, EE**	**E**	**s**	**B**	**R, CS, D/C, CR**	**-**
Interdisciplinary Scenarios	(Soobard & Rannikmäe, 2011)	**OP**	**EE, OT**	**S**	**n/u**	**ES**	**R, CT, CR**	**-**
IQB state comparison	(Pant et al., 2013)	**MI**	**Q, HG, EG, EE, OT**	**S**	**s**	**B, C, P**	**D/C**	**+**
National Assessment of Educational Progress (NAEP) Science Assessment	(National Assessment Governing Board, 2007; United States National Assessment Governing Board et al., 2010)	**MI**	**EG, EE, DC, OT**	**E, S**	**s**	**B, ES, P**	**na**	**+**

Test name	References	Test format[a]	Covered scientific reasoning skills[b]	Target group(s)[c]	Assumption about domain generality[d]	Context domain(s)[e]	Checks of psychometric properties[f]	Test norms[g]
National Assessment Program - Science literacy	(Donovan, Hutton, Lennon, O'Connor, & Morrissey, 2008; Donovan, Lennon, O'Connor, & Morrissey, 2008; Wu, Donovan, Hutton, & Lennon, 2008)	**MI**	**Q, HG, EG, EE, DC, CS, OT**	**E**	**n/u**	**B, C, ES, P**	**R, CS**	**+**
Natural Sciences Methods Test (NAW)	(Klos, 2009; Klos, Henke, Kieren, Walpuski, & Sumfleth, 2008)	**MI**	**HG, EG, DC**	**S**	**n/u**	**C**	**R, CS, D/C, CR**	**-**
Objective Referenced Evaluation in Science (ORES)	(Shaw, 1983)	**MC**	**HG, EG, EE, DC**	**E**	**n/u**	**na**	**R, CT, CS, CR**	**-**
Online Portfolio Assessment and Diagnosis Scheme (OPASS)	(Su, Lin, Tseng, & Lu, 2011)	**AA**	**HG, EG, DC, CS**	**S**	**n/u**	**B, P**	**CT, D/C, CR**	**-**
PISA science 2006	(OECD, 2006, 2007, 2009)	**MI**	**PI, HG, EG, DC, CS, OT**	**S**	**n/u**	**NS**	**R, CT, CS, D/C**	**+**
Practical Tests Assessment Inventory (PTAI)	(Tamir, Nussinovitz, & Friedler, 1982)	**SC**	**PI, HG, EG, EE, DC, CS, OT**	**S**	**n/u**	**B**	**CT, CS**	**-**
Processes of Biological Investigations Test (PBIT)	(Germann, 1989)	**MC**	**HG, EE, DC**	**S**	**s**	**B**	**R, CS, D/C, CR**	**-**
Research Knowledge Skills to Conduct Research Questionnaire	(Meerah et al., 2012)	**SA**	**EG, CS, OT**	**U**	**n/u**	**na**	**R, CT, D/C**	**-**
Rubric	(Timmerman, Strickland, Johnson, & Payne, 2011)	**SC**	**HG, EG, EE, DC, CS, OT**	**U**	**n/u**	**B**	**CT, CR**	**-**

Test name	References	Test format[a]	Covered scientific reasoning skills[b]	Target group(s)[c]	Assumption about domain generality[d]	Context domain(s)[e]	Checks of psychometric properties[f]	Test norms[g]
Science Process Skill Test (SPST)	(Feyzioglu et al., 2012)	**MC**	**HG, EG, EE, DC**	**S**	**n/u**	**C**	**R, CT, CS, CR**	**-**
Science-P	(D. Mayer, 2012; D. Mayer, Sodian, Koerber, & Schwippert, 2014)	**MI**	**EG, EE, OT**	**E**	**g**	**na**	**R, CS, D/C, CR**	**-**
Scientific Reasoning Test, Version 9 (SR-9)	(Sundre, 2008)	**MC**	**HG, EG, OT**	**U**	**g**	**na**	**R, CT**	**+**
Springs task	(Linn & Rice, 1979)	**OT**	**EG, CS**	**E, S, U**	**g**	**na**	**R, D/C**	**-**
Test of competencies of scientific thinking	(Grube, 2010)	**OP**	**Q, HG, EG, EE**	**E**	**s**	**B**	**R, CT, CS, CR**	**-**
Test Of Enquiry Skills (TOES)	(Fraser, 1979, 1980)	**MC**	**EG, EE, DC, CS, OT**	**E, S**	**n/u**	**NS**	**R, CS**	**-**
Test of Integrated Process Skills (TIPS) I&II	(Baird, 1989; Burns, Okey, & Wise, 1985; Dillashaw & Okey, 1980)	**MC**	**HG, EG, EE**	**S**	**g**	**na**	**R, CT, CS, CR**	**-**
Test Of Logical Thinking (TOLT)	(Tobin & Capie, 1981)	**MC**	**EG, EE, OT**	**E, S, U**	**g**	**na**	**R, CS, D/C, CR**	**-**
Test of Science Process Skills	(Molitor & George, 1976)	**MC**	**EE, DC**	**E**	**g**	**na**	**R, CS, D/C, CR**	**-**
Test Of Scientific Literacy Skills (TOSLS)	(Gormally, Brickman, & Lutz, 2012)	**MC**	**EG, EE, DC, CS, OT**	**U**	**g**	**na**	**R, CT, CS**	**-**

Test name	References	Test format[a]	Covered scientific reasoning skills[b]	Target group(s)[c]	Assumption about domain generality[d]	Context domain(s)[e]	Checks of psychometric properties[f]	Test norms[g]
Test of Scientific Thinking (TST)	(Frederiksen & Ward, 1978; Ward, Frederiksen, & Carlson, 1980)	**OP**	**HG, EG, CS**	**U**	**s**	**BS**	**R, D/C, CR**	**-**
TIMSS	(Martin & Mullis, 2012; Martin, Mullis, Foy, & Stanco, 2012; Mullis, Martin, Ruddock, O'Sullivan, & Preuschoff, 2009)	**MI**	**Q, HG, EG, EE, DC, CS**	**E, S**	**s**	**B, C, ES, P**	**R, CS, D/C**	**+**

[a]Test format: "AA" automated analyses of simulated experiments, "MC" multiple-choice questions, "MI" mixed question format, "OP" open-ended questions, "OT" other question format, "SA" self-assessments, "SC" scoring rubrics for reports about conducted experiments. [b]Scientific reasoning skills: "PI" problem identification, "Q" questioning, "HG" hypothesis generation, "EG" evidence generation, "EE" evidence evaluation, "DC" drawing conclusions, "CS" communicating and scrutinizing, "OT" other skill. [c]Target group(s): "E" elementary school students, "S" secondary school students, "U" university students. [d]Domain generality assumptions: "g" test assumed to be domain-general, "s" test assumed to be domain-specific, "g/s" different assumptions for different parts, "n/u" assumption not stated or unclear. [e]Context domain(s): "B" biology (including life sciences), "BS" behavioral sciences and psychology, "C" chemistry (including natural and processed materials), "ES" earth and space science (including geography), "M" medicine, "NS" natural sciences (no specification), "P" physics (including energy and change), "na" context domain not (clearly) given. [f]The following checks of psychometric properties were reported: "R" reliability, "CT" content validity, "CS" construct validity (other than divergent or convergent validity), "D/C" divergent and/or convergent validity, "CR" criterion validity. [g]Test norms (criterion or population based): "+" norms are reported; "-" no norms reported.

2.3.2 Core skills addressed by the tests

Most tests focused on three to four skills to assess scientific reasoning (M = 3.39). *Evidence generation* was the most frequently included skill in scientific reasoning tests; 33 of 38 tests had some form of assessment of this skill. Other prioritized skills included in at least half of the test instruments were *hypothesis generation*, *evidence evaluation*, and *drawing conclusions* (see Figure 1). This pattern held true for older and newer tests alike. There was one skill that was included in newer tests but not in older tests: *questioning*. Of the 258 sorted skill descriptions, 40 were sorted into the "other" category. The main reason a skill description did not fit into the coding scheme and had to be sorted into the "other" category was that some tests did not only measure scientific reasoning skills but also knowledge or understanding the nature of science. Since all skill descriptions of included tests were sorted, some of these other skills got into the pool of descriptions. Apart from these instances, skill descriptions that did not fit into the eight main categories of the coding scheme and thus had to be sorted into the "other" category were referring to quantitative skills (nine descriptions) or to societal or ethical issues of science (five descriptions).

2.3.3 Theoretical background and dimensionality

When it comes to scientific reasoning conceptualizations used for test construction, 15 test authors stated that they had used a specific theory (one of them referred to two theories). The first category of theories – they assume one general ability developing over time (e.g. Inhelder & Piaget, 1958) – was used by five tests. The second category of theories – postulating several independent skills (e.g. Livermore, 1964) – was used by four tests. The third and most common alternative among more recent tests is to assume multiple skills but to conceptualize them as being part of a problem-solving process (e.g. Klahr & Dunbar, 1988). This last kind of theory was used by seven tests, all of them from the second wave of

test development (2002 – 2013). During this second wave the first and second type of theories were only used two times and one time, respectively. Thus, considering the two waves of test development, there seems to have been a shift from assuming scientific reasoning to be a unidimensional ability to considering scientific reasoning to be a problem-solving activity in which several skills have to be orchestrated.

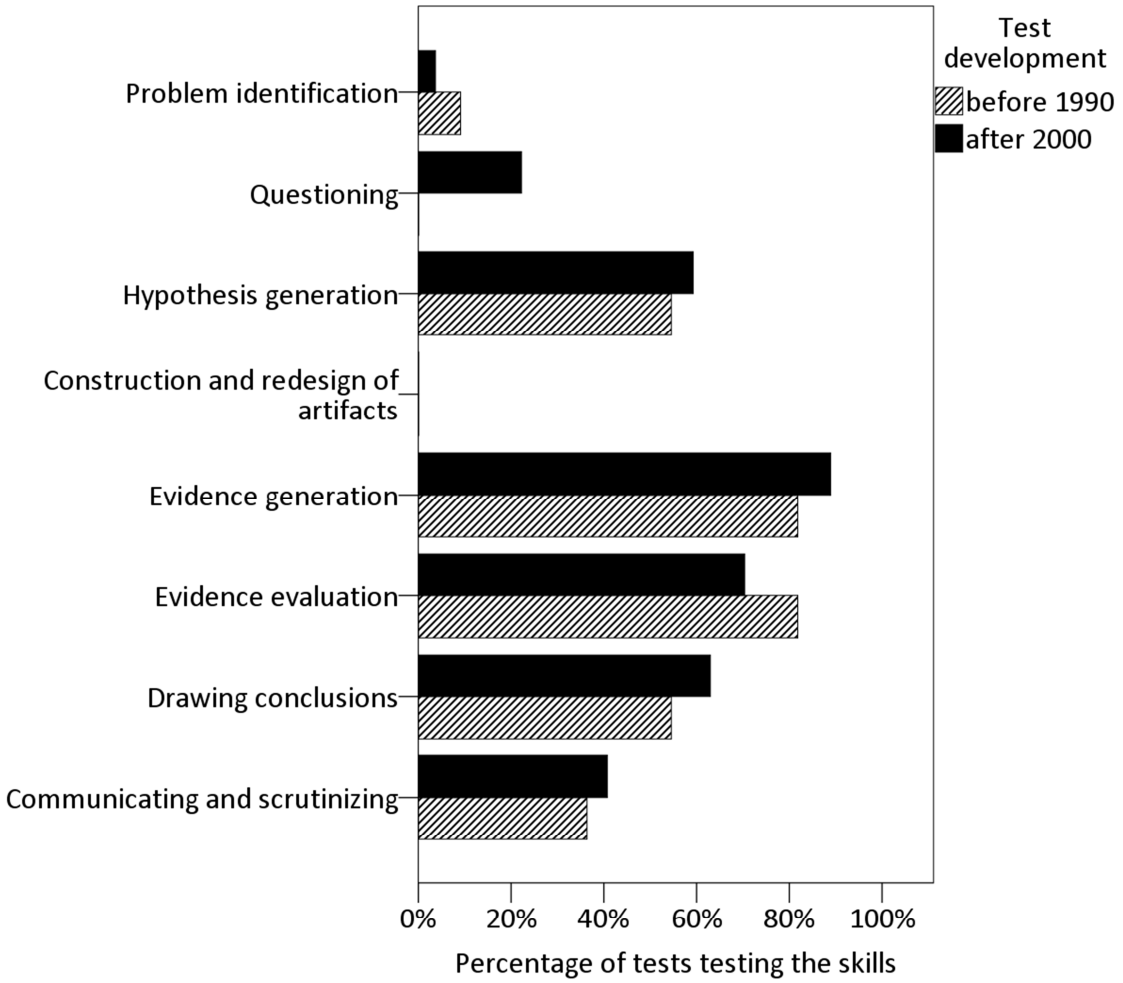

Figure 1. Scientific reasoning skills covered by the tests sorted by year of development.

A total of 15 tests used educational standards as a basis for test construction. Especially during the second wave of test development this became more common, with 13 tests choosing this path. Prime examples for this trend were large-scale assessments. They take an

interesting middle position in the controversy of a single versus multiple dimensions. Large-scale assessments typically assume various different skills but also one underlying factor that unites them. Often, this single uniting factor is a major focus in the report about results. For instance, the PISA framework (OECD, 2006) differentiates between the three skills *identifying scientific issues*, *explaining phenomena scientifically*, and *using scientific evidence* but also combines their scores into a single science scale.

In theory, results from model tests could help to answer the question of dimensionality. However, results from model tests were too few in number and differed too much to give a general answer of how well the tests fitted the assumed theoretical structure and thus we cannot give a decisive empirical answer on the question of dimensionality.

2.3.4 Test context and assumptions about domain generality vs. specificity

Not surprisingly, most test instruments for scientific reasoning used specific science domain contexts; biology was most common (k = 13), followed by chemistry and physics (k = 8 for both domain contexts), earth and space science (k = 5), and, less frequently, natural science (without specification; k = 2) medicine (k = 2) and social sciences (k = 2). It should be noted that our judgment of the text context allowed tests to be embedded in more than one or no domain context at all. The use of a specific domain context did not necessarily imply that the test authors assumed that scientific reasoning is specific for this particular domain.

Tests such as the often-used Classroom Test of Scientific Reasoning (Lawson, 1978) assume that scientific reasoning is distinct from specific domain knowledge and testable in a generally valid way. However, other tests – such as the Chemistry Concept Reasoning Test (Cloonan & Hutchinson, 2011) – assume domain specificity. Overall, about one-third of all test authors were categorized as assuming domain generality (k = 12), one-third were categorized as assuming domain specificity (k = 12), and the remaining third were

categorized as not providing clear assumptions about specificity or generality (k = 13). One test was categorized as making different assumptions about different parts of the test (Brown et al., 2010). In comparing the first and the second wave of test development, there was a trend away from generality and towards specificity assumptions. Of the 12 tests that assume domain specificity, ten were from the second wave. However, making assumptions about domain specificity and testing them are two different things. Only four test authors tested their assumptions on domain generality or specificity. There was one test checking its generality assumption and it was successful in doing so. Two of three tests assuming specificity were successful with their test of this assumption.

2.3.5 Psychometric properties and norms

Apart from the issues mentioned in the last two sections about the lack of checks of conceptual assumptions, there were some other noteworthy points regarding the psychometric properties of tests. The number of tests checking their reliability decreased for newer tests. In the first wave 10 out of 11 tests reported reliability checks but only 17 out of 27 newer tests did so. Regarding validity checks, there were only six tests using other scientific reasoning tests to establish convergent validity and only four tests that used measures of general cognitive abilities like IQ tests to establish divergent validity. We only found one test that used a longitudinal approach and tried to establish predictive validity. The test results were correlated with results from a questionnaire (covering, for instance, the selected graduate program, professional preferences, a self-evaluation of knowledge and skills and successes in the first year of graduate school) that was given to participants 1.5 years later (Frederiksen & Ward, 1978). The number of significant correlations between test results and these indicators of the quality of science careers was just barely above the chance level of 5%, indicating that

there were probably no meaningful connections. Last, we discovered that criterion or population based norms existed for seven tests.

2.3.6 Approaches to the measurement of scientific reasoning

The test instruments used a broad variety of test formats that fell on the following continuum: From closed tests, in which test takers have to answer questions about material given to them, to open test formats, in which students have to produce something on their own. Within the former, we found multiple-choice tests, such as the Classroom Test of Scientific Reasoning (Lawson, 1978), in which test takers see a diagram showing different weights attached to strings with different lengths and then have to answer two multiple-choice questions. The first question asks which strings should be used to find out whether the length of the string has an influence on the time to swing back and forth, and the second requires that the student indicates the explanation for the answer.

Overall, 14 of the 38 tests were purely multiple-choice but test developers have used alternative formats more often since the early 2000s in particular. Still more on the closed side of the test format spectrum were tests ($k = 2$) that let test takers rate their own skill level in regard to different scientific reasoning skills like choosing suitable study methods or recording data (e.g. H.-P. Chang et al., 2011) Some tests ($k = 9$) like PISA (OECD, 2006) add open-ended questions to their mix which in some cases still aim for a very particular answer. On the middle ground of the closed–open continuum, there were recent tests ($k = 2$; e.g. Gobert et al., 2013) using simulated experiments that are analyzed automatically. Several drop-down menus ensure that students can build their own hypotheses but stay within a set of given options. They can set the parameters for an experimental design and see the simulated results. An algorithm automatically analyses if students were able to design controlled experiments. Finally, at the open-ended side, there were test formats ($k = 2$) like the Rubric

(Timmerman et al., 2011) which provides a standardized scheme to analyze biology lab reports of students, helping to evaluate (amongst others) if the hypotheses are stated clearly, data is analyzed properly, and conclusions are drawn logically. The remaining eight open tests ask, for instance, for the description of an experiment to test a claim (e.g. a new iron supplement will improve memory; Sirum & Humburg, 2011), and are rated according to the criteria the test taker addresses (e.g. correctly determining the independent and dependent variables).

2.4 Discussion

Since the beginning of the millennium there is a resurgent interest in the measurement of scientific reasoning that coincides with a set of new educational standards (NRC, 1996) and results from large-scale assessments like PISA (OECD, 2006). What scientific reasoning entails and how it is conceptualized and measured has clearly evolved over these last two decades according to the 38 scientific reasoning tests we reviewed in this article. There is a shift away from considering scientific reasoning as having one single underlying cognitive ability developing in childhood and youth and that is used for scientific reasoning in any domain. Instead, there is a trend towards conceptualizing the competence as a domain-specific set of different but coordinated skills. Consequently, more recent tests assume a multidimensional structure of the scientific reasoning construct. The addition of *questioning* to some newer test might be reflective of a trend towards model based inquiry in which questions are not just handed to students (Windschitl et al., 2008). However, there is still the same number of core scientific reasoning skills (e.g. *evidence evaluation*) that are included in most tests and this number has hardly increased with the shift from uni- to multidimensional models of scientific reasoning. Skills referring to quantitative reasoning were more often

included than e.g. *questioning* or *problem identification*. Apparently, at least some authors see quantitative reasoning as a relevant aspect of scientific reasoning itself, instead of being an overarching competence (Shavelson & Huang, 2003), so in the future there should be a discussion within the field if and in which way quantitative reasoning should be part of the conceptualization of scientific reasoning. Although the number of assessed skills does not increase in recent years, the test formats become more diverse. Multiple-choice tests are not as common as they used to be. Instead, new test formats using virtual experiments begin to appear. This might be reflective of the shift in science education towards practicing science in addition to teaching knowledge about phenomena and research procedures (NRC, 2012).

Clearly, several challenges still remain: Hardly any tests exist that aim to assess scientific reasoning skills in the general population outside of formal education institutions. Assumptions about dimensionality and domain generality are rarely psychometrically tested. Regarding the topic of domain generality there exists an additional challenge, namely that researchers rarely tap into the questions of whether different skills might have a different degree of generality and/or if some skills would transfer to some domains but not to others. For instance, it seems plausible that test items about developing a valid research question are usable in empirical as well as non-empirical domains but items asking for the experimental generation of evidence only apply to domains that work empirically.

Authors should also compare results of scientific reasoning tests to the results of other scientific reasoning tests more frequently to find out more about the homogeneity of different measures of scientific reasoning skills. In order to embed scientific reasoning into a nomological network and thus to better understand what scientific reasoning does and does not entail, more focus should be placed on the differences between scientific reasoning tests and tests that are intended to measure something else, particularly other cognitive constructs (e.g. intelligence). Besides, if we cannot separate scientific reasoning from other cognitive

constructs it will be hard to justify the time and effort spent on creating scientific reasoning assessments. The multitude of criterion validity measures makes it hard to establish common standards for what scientific reasoning test results should be able to predict, i.e. the relevance of the test results. In particular, we need to know more about the role of scientific reasoning in predicting long-term effects with respect to learning, academic achievement, and understanding scientific studies. Since different test formats exist, it might be interesting to compare them against each other and see if different test situations call for the use of different formats and if, and in regard to which aspects, newer test formats are superior to older multiple-choice tests.

These shortcomings might be the reason that so far it is not common to see a scientific reasoning test as an outcome measure of a training in scientific reasoning. The limitations might serve as an excuse to develop some measure with unknown psychometric properties that has a high chance of showing that the intervention works, because it is still relatively easy to argue that existing scientific reasoning instruments are not superior to such an approach. If that would be the case it would be even more important to close our knowledge gaps about scientific reasoning tests. Only if we have instruments with well-known psychometric properties we can demand that different interventions should be compared with the same measure to compare their effectiveness. Consequently, we need to know more about the structure and psychometric properties of our current measures. However, it should be mentioned that there is a simpler explanation for the rare use of the tests. It might be that the research community is just not aware of existing tests. If that is the case, this review is a contribution to solving this problem.

While hopefully giving some useful insights, probably the biggest limitation of this review is that we had to make a selection out of all the skills mentioned somewhere in the many different conceptualizations of what makes a scientifically literate person. Readers who

were mainly looking for tests of *nature of science* or argumentation will not be satisfied with the selection presented in this review. At least in the field of *nature of science* there seems to be a small number of already established scales and hence less need for an overview. In comparison to the skills covered in this review, the Views of Nature of Science Questionnaire (VNOS) by Lederman et al. (2002), a typical *nature of science* assessment, asks questions like "is there a difference between scientific knowledge and opinion?" and thus rather covers knowledge about science than a scientific reasoning skill.

Considering the present state of the field, what might be best practice recommendations for people in need of a scientific reasoning test? Keeping in mind the psychometric limitations we mentioned, the missing knowledge about the predictive power for later academic and scientific performance in particular, we would advise against basing high-stake decisions, especially about individuals, on current scientific reasoning tests. However, we do think that some valuable insights can be gained from scientific reasoning tests, especially on the group level, such as an entire science class. Here, one of several tests can be used by practitioners to inform teaching and by researchers for determining the effects of an intervention in an experimental setting. For instance, if a university teacher from the social sciences wants to know if a class about how to construct a good experiment had an effect, tests like the EDAT (Sirum & Humburg, 2011) or the CISRS (Weld et al., 2011), in which test takers have to describe a way to test a claim, should provide some useful answers.

In light of the findings from this review, we suggest the following pragmatic approach for practitioners and researchers who want to use a scientific reasoning test: Instead of developing a new test, researchers should start with using the list of scientific reasoning tests in Table 3. By considering the constraints of a concrete assessment situation like the domain, the skills that the test should cover, desired test formats, or the age of the target group it is probably straight forward to identify a small number of promising tests. For instance, if the

test should target university students, assume domain generality of scientific reasoning and use a multiple-choice format, there are four potential tests in the list of scientific reasoning tests (Gormally et al., 2012; Lawson, 1978; Sundre, 2008; Tobin & Capie, 1981). Finally, inspect these candidate tests in order to select the one with the best fit to the intended purpose. The first priority of the inspection should be to make sure that the scientific reasoning conceptualization that is used by the test matches with the construct that the practitioner or researcher is interested in. Of course, the results from the checks of psychometric properties should also play an important role in the decision.

However, this approach can only serve as a temporary solution. With the increasing emphasis of scientific reasoning as a process and as a desired outcome of education, we need to consider the assessment of scientific reasoning more systematically. Although there have been some developments in testing scientific reasoning, the quality of measurement is still largely unclear. As a strategy for future research we suggest the following: Focus on testing assumptions about the structure and domain generality or specificity, find out more about the relevance of scientific reasoning test results (especially regarding the aforementioned long-term effects on learning, academic achievement and understanding scientific studies), and compare different scientific reasoning tests with each other (to find out more about which tests are better in general or for specific purposes). A lot of knowledge can be gained about these three issues with existing tests and we suggest that new tests should only be developed when they also contribute to resolving these issues. In the case of new tests being developed, another aim should be to include all the skills that are deemed relevant in the scientific reasoning conceptualization that is used, not least because otherwise it is possible that only the skills that are on the test will be taught.

3 Study 2 – A Detailed Analysis of a Common Scientific Reasoning Test

3.1 Introduction

The assessment of competencies in higher education is gaining in importance and several new tests have been developed but overall there is still a big research gap in this area (Zlatkin-Troitschanskaia, Shavelson, & Kuhn, 2015). One central aspect of higher education competencies are scientific reasoning skills. Students have to be able to define problems, formulate questions and hypotheses, gather and evaluate evidence, and explain and communicate results (F. Fischer et al., 2014; NRC, 2012). Some conceptualizations of scientific reasoning also include quantitative reasoning (NRC, 2012) while others do not (F. Fischer et al., 2014).

While scientific reasoning skills should definitely be taught throughout all levels of education, they are especially relevant at a university level: Scientific reasoning skills are important for the workforce (The Royal Society, 2014), and university is the last stage of formal education for many students before they will work in jobs that rely heavily on these skills. Thus, it is crucial to know how proficient the scientific reasoning skills of our current students are. Besides, the majority of students will never again be so close to on-going cutting edge research as they are in university, so higher education should be an excellent learning opportunity for scientific reasoning skills.

Several scientific reasoning tests exist that can be used with university students but their quality is in large parts still unknown (Opitz, Heene, & Fischer, 2015). This study addresses the following four aspects – which will be explained in detail in the next

paragraphs – of assessing scientific reasoning skills: the debate around domain generality vs. domain specificity, divergent validity, factorial structure, and criterion validity. To do so, this study analyzed an often used scientific reasoning test, the Lawson Classroom Test of Scientific Reasoning (CTSR; Lawson, 1978), in great detail employing a range of statistical methods, some of which have only been recently developed. The results are used to learn more about the test at hand, future test construction and evaluation, and the construct of scientific reasoning itself.

3.1.1 Domain generality vs. domain specificity of scientific reasoning

While some test authors assume that their test measures scientific reasoning in a domain-general way, others assume that their scientific reasoning test is domain-specific (Opitz et al., 2015). The extent to which these assumptions are actually tested is the subject of the next paragraph. However, first of all we have to explain necessary terms and state why this debate is relevant: Domain generality implies that scientific reasoning skills are the same across domains such as physics or biology and thus can be assessed independent of the context the skill is tested in. Domain specificity implies that scientific reasoning skills are so closely tied to a domain that a test can only ever measure scientific reasoning within one domain and thus the result cannot be transferred to another context. This debate becomes especially relevant on the university level because students are normally divided into majors and get influenced by one subject area. If one test is supposed to be able to measure the scientific reasoning skills of all students – as a domain-general test would assume – we first have to be sure that students have the same chance to perform well in the test independent of the subject area they are studying. The only factor that should matter is the scientific reasoning skill of the test taker, not what major they are enrolled in. This does not mean that there cannot be significant differences between students from different majors regarding the

level of their scientific reasoning skills. It just means that if the test has items that are embedded into a certain context, e.g. biology, but that are supposed to measure scientific reasoning in a domain-general way, biology students should not have an unfair advantage in these items.

So far, the assumptions about the domain generality or domain specificity of assessments have hardly been tested at all. One of the few exceptions is the study by Weld et al. (2011) who checked their domain generality assumption by showing that there was no significant difference between biology majors and elementary education majors. Similarly, Cloonan and Hutchinson (2011) tested their assumption of domain specificity by comparing the mean score achieved by chemistry graduate students with the score of people who hold a high degree of education outside of chemistry. However, this method of testing a domain related assumption is problematic because it confounds two aspects that should be separated, namely the observed test score and the underlying latent ability. For instance, in the aforementioned study that tested domain generality (Weld et al., 2011) it is possible that the two groups only achieved a similar total score because one group had an unfair advantage that canceled out the fact that their actual skill level was lower compared to the other group. Similarly, if the biology majors would have had a significantly higher score this could still mean one of two things: Either they simply have better scientific reasoning skills, which would not run counter to the domain generality of the test, or the test contained elements that gave biology students an unfair advantage, which would run counter to the domain generality of the test. A simple mean difference cannot tell these scenarios apart from one another. Besides, the absence of bias from the level of total scores does not mean that there is no bias on the item level (Borsboom, 2006). If we want to learn more about constructing good scientific reasoning tests we have to be aware of bias on the item level and should not be satisfied with checking bias only on the level of total scores.

This study will compare currently used methods to establish domain generality based on classical test theory (CTT), e.g. item difficulties and mean differences, and newer methods based on item-response theory (IRT). The central IRT concept in this regard is differential item functioning (DIF). DIF can be used to establish the measurement invariance of a test. If measurement invariance holds, the test has the same factorial structure in two groups. In contrast, a violation of invariance indicates that the test results are influenced by group membership and therefore a comparison of the two groups will be biased. Characteristics that are commonly used in evaluations of DIF are gender, minority status, race, sub-culture, or language (Zumbo, 1999). For instance, a DIF test has been used to search for language bias in a scientific inquiry test (Turkan & Liu, 2012). The novel idea of this study is that domain generality can be framed as a problem of measurement invariance, too, and that it can be evaluated by exploring DIF. The underlying assumption is that if a test cannot even show that it has the same factorial structure for students from two different domains, it cannot possibly claim to measure scientific reasoning in a domain-general way.

IRT based methods have been proven to be effective in establishing measurement invariance. IRT based methods are more sensitive to differences than methods based on confirmatory factor analysis (Reise, Widaman, & Pugh, 1993). They are also especially suited for discovering DIF on an item level (Molenaar & Borsboom, 2013) and as pointed out before, this is a very desirable property when evaluating scientific reasoning tests. While past methods to check DIF could only compare two groups at a time, newer methods overcame this problem. This study will look at two of these newer methods and compare their results in order to see how stable the results are across different methods.

The first method is model-based recursive partitioning, better known as tree models (Strobl, Malley, & Tutz, 2009). Within the class of tree based analyses we will employ two techniques. One technique, a so-called Rasch tree, will be used for establishing DIF on the

level of the whole test (Strobl, Kopf, & Zeileis, 2015). The other technique, item-focused trees, is used subsequently to evaluate DIF on the item level (Tutz & Berger, 2016). The second method utilizes a so-called lasso (least absolute shrinkage and selection operator) technique (Tibshirani, 1996). The concrete technique that will be used was developed by Tutz and Schauberger (2015). These techniques were selected because all of them have demonstrated that they are capable of detecting DIF, that they have an especially low rate of false positive indications of DIF (which is very important in a situation like ours in which many parameters are estimated), and that they result in interpretable models (Strobl et al., 2015; Tutz & Berger, 2016; Tutz & Schauberger, 2015). However, it should be noted that the lasso based method leads to the underestimation of the absolute value of the bias due to the shrinkage of all parameters (Tutz & Schauberger, 2015).

In case we would discover DIF we were also interested in observing whether there is a relation between DIF and the level of domain dependency of an item. To establish a coding scheme for this domain dependency we drew inspiration from a conceptualization by Shavelson and Huang (2003). The conceptualization assumes different levels of broadness for cognitive abilities. The range reaches from general intelligence on the highest level to abilities such as verbal reasoning, quantitative reasoning and spatial reasoning on a high intermediary level, to abilities that are relevant in broad domains like science on a low intermediary level, and finally to knowledge that is required in a domain on the lowest level. We expected that there is a higher chance for a domain-specific advantage if lower levels play a role in solving an item.

3.1.2 The connection of scientific reasoning with general reasoning and science knowledge

The model by Shavelson and Huang (2003) also points to another challenge for scientific reasoning: Is the construct distinct from general reasoning and science knowledge? It is necessary to know what a test exactly measures if we want to know what the performance in a test implies about a student (Pellegrino, 2013).

If general reasoning can account for the variance in scientific reasoning tests, as some claim (Simon, 1966), then it would be hard to justify the effort of assessing scientific reasoning in addition to general reasoning. Several studies have dealt with this distinction: Schunn and Anderson (1999) argued that skills such as designing experiments and interpreting results cannot be the same as general reasoning because otherwise undergrads would have been as good as expert researchers at testing claims in a novel field. In terms of scientific reasoning assessment, one study showed that there is a significant correlation between scientific reasoning and figural reasoning in a secondary school setting, but even when taken together with other individual factors this did not explain more than 20% of the variance (Pant et al., 2013). Another test for middle schoolers found correlations of .36 and .45 with a complete general reasoning scale in seventh grade and 12th grade, respectively, and a correlation of .26 with just the verbal subscale in the seventh grade (Klos et al., 2008). In an analysis of a scientific reasoning test for elementary school students a model with the two dimensions intelligence and scientific reasoning achieved a better fit compared to a model where both constructs were combined into one dimension (D. Mayer et al., 2014). The correlation of the two latent dimensions was .62 while the manifest correlation was .38. In a similar analysis of the same test a model with three dimensions (intelligence, verbal reasoning, and scientific thinking) achieved a better fit compared to a model with one combined dimension (Koerber, Mayer, Osterhaus, Schwippert, & Sodian, 2015). This time,

the latent correlation between intelligence and scientific reasoning was .63. So far, no results exist for tests that aim at university students.

This study makes a first step towards closing this gap. On top of manifest and latent regressions this study also uses a bifactor model to explore the relation between scientific reasoning and general reasoning. The idea is to treat scientific reasoning as a subscale of general reasoning. When we treat scientific reasoning this way despite the fact that scientific reasoning is not as closely related to general reasoning as other reasoning scales, scientific reasoning should stand out as more independent from a general factor in comparison to other reasoning scales. Reise, Bonifay, and Haviland (2013) described ways to test the independence of subscales from a general factor. One of the criterions that will be used in this study is the so called Haberman criterion (Haberman, 2007). It evaluates whether observed subscale scores are better than the total score in predicting true subscale scores. The Haberman criterion is meant as a minimum hurdle that should always be followed up by other checks if it is met. Further calculations suggested by Reise et al. (2013), which are described in the Methods section, served as this follow-up in our study.

This study is also going to explore the strength of the connection of scientific reasoning with science knowledge. Two studies have shown that there is a significant relation between the CTSR and science knowledge (Lawson, Clark, et al., 2000; Lawson, Alkhoury, Benford, Clark, & Falconer, 2000). This study will add to this knowledge base.

3.1.3 The factorial structure of scientific reasoning

Scientific reasoning conceptualizations differ in their assumptions about the dimensionality of the construct (Opitz et al., 2015). Many tests calculate a single sum score which implies a single underlying skill, but we do not know enough to judge if this procedure is actually justified. The answer to the question of whether scientific reasoning is a single

latent skill with exchangeable facets or rather just a descriptive term for several independent skills is important for how specific we have to train scientific reasoning as well as for test construction and analysis. For instance, if scientific reasoning consists of several independent skills we need separate subscales or tests for all of them and cannot simply calculate a single sum score and expect that it reflects scientific reasoning accurately.

The available evidence about the dimensionality of scientific reasoning is inconclusive. The CTSR claims to be unidimensional despite the fact that an exploratory factor analysis of its original version resulted in a three-factor model (Lawson, 1978). Items about displaced volume and the conservation of weight both formed separate factors. Pratt and Hacker (1984) supported the argument for a multidimensional structure with an IRT based analysis. In a study with 150 students they tested the unidimensionality of the CTSR with a Rasch model and found three misfitting items. One of these items, an item about pouring water from one cylinder into another one, is still included in the most recent version of the test. To our knowledge, there are no studies about the factorial structure of the CTSR that use its most recent version.

Studies that analyzed the factorial structure of other scientific reasoning tests came to a whole range of different conclusions. Looking at the results from various factor analyses it is possible to find not only one-factor models, but also two-factor models, three-factor models, four-factor models, five-factor models, eight-factor models, and 11-factor models (H.-P. Chang et al., 2011; Feyzioglu et al., 2012; Germann, 1989; Gormally et al., 2012; Grube, 2010; Hammann, Phan, & Bayrhuber, 2008; Klos et al., 2008; Nowak et al., 2013; Roberts & Gott, 2004; Tobin & Capie, 1981). A particularly interesting study was conducted by Li et al. (2006). They showed that a three-factor model, which split a scientific reasoning test into three knowledge components (declarative, procedural, and schematic knowledge) was a better fit than several other models that made different splits.

3.1.4 Criterion validity

While some see teaching scientific reasoning as an end to itself (Harlen, 1999) everyone else should be concerned about the criterion validity of scientific reasoning tests. The main problem in terms of criterion validity of scientific reasoning tests is that almost all tests that tried to establish criterion validity chose a different criterion which makes it hard to draw any meaningful conclusions (Opitz et al., 2015). This study will employ the most commonly used criterion – grades – by evaluating the relation between scientific reasoning and students' grade point average (GPA) in their undergraduate studies as well as bachelor thesis grades. Evaluating this relation can give us hints about the connection of scientific reasoning and academic achievement. Additionally, for medical students, this study will explore the relation of scientific reasoning and the ability to detect errors in a diagnosis, which is a relevant skill for practicing physicians.

3.1.5 Goals and questions

The goal of this study was not only to find out more about a commonly used scientific reasoning test: By conducting detailed analyses, some of them on the item level, of the domain generality, the divergent validity, the factorial structure, and the criterion validity of the test, we also wanted to learn more about scientific reasoning assessment and the construct of scientific reasoning in general.

These were the questions we wanted to answer: First, is the overall test as well as the single items measurement invariant? If that is not the case the basic requirement for domain generality is not fulfilled. Second, if DIF is found, is there a relation between the bias and the domain dependency of the items? This is an important follow-up investigation to find out more about the nature of the potential domain bias. Third, do IRT methods provide us with information that is relevant to evaluate the domain generality assumption and that goes

beyond the information acquired by the methods of CTT? Only an added informational value would justify the extra effort that these analyses entail. Fourth, are there differences in the relation between scientific reasoning and general reasoning as well as science knowledge between different majors? If so, this would be another indicator of domain specificity. Fifth, what is the relation between scientific reasoning and general reasoning as well as science knowledge when we look at the complete sample? Sixth, if we treat scientific reasoning as a subscale of general reasoning will it stand out as more independent than other indicators of general reasoning? The answers to the fifth and sixth point will contribute to the question of how closely connected scientific and general reasoning are. Seventh, what is the factorial structure of the CTSR and scientific reasoning in general? This will be relevant for creating more accurate conceptualizations and it will inform us about how to construct tests in the future. Eighth, what is the criterion validity of the CTSR? A high criterion validity would support the relevance of assessing scientific reasoning.

The study we conducted to answer these questions was set in a higher education setting. Considering that there is an especially big research gap in the assessment of cognitive competencies at this educational level, this was an appropriate choice.

3.2 Methods

3.2.1 Sample

Our aim was to collect data from at least 500 students, because simulation studies have shown that the intended DIF analyses have an acceptable hit rate and will produce few false alarms in samples of this size (Strobl et al., 2015; Tutz & Berger, 2016; Tutz & Schauberger, 2015). The participants had to be university students majoring in physics, biology, or medicine. We chose physics and biology because the CTSR contains both items with a

biology context and items with a physics context, so we needed to test students from the according majors to explore domain-specific advantages. We chose medicine as the comparison domain, because from an epistemological perspective it is not too distant from the other two domains. This allowed us to check if there is a domain-specific advantage for all students with natural science majors but not for other students even if they study closely related subjects. The students had to be at the undergraduate or graduate level, so PhD students were not allowed to participate. Additionally, the students had to be enrolled in at least their fifth semester because we reasoned that if domain effects exist they will not be apparent at the start of higher education. It should be noted that in Germany students are starting their major directly in the first semester. Depending on the major, some students will still attend classes that cover content areas from other majors, which is especially true for medical students. They will learn both about physics and biology. Lastly, participants had to be fluent in German.

We excluded participants who did not reach the end of the test booklet. We had to exclude 27 students from the online test because of this. Furthermore, we excluded online test participants when they had spent less than 40% of the maximum time on the CTSR or the scientific knowledge test (in a pilot phase this seemed to be the lowest realistic boundary for finishing the test) and answered at chance level or worse, indicating that they were just randomly answering the questions. We had to exclude one student for this reason.

The complete sample consisted of 507 students. Some participants reached the end of the test booklet but did not answer all tests and for some online participants it could not be verified if they stayed within the maximum time limits for all of the tests. Therefore, the sample size for the single tests is below the total number of participants. Not counting tests that were left out by participants and tests with an unverified completion time, the scientific reasoning test was completed by 506 students, the figural reasoning test by 496 students, the

verbal reasoning test by 489 students, the numerical reasoning test by 492 students, and the science knowledge test by 502 students.

The demographic variables were as follows: 249 students were male, 256 were female, and 2 participants did not answer this question. The age of participants was $M = 23.01$ years ($SD = 2.91$) and they had studied $M = 6.76$ semester ($SD = 2.53$). The most common major was physics (192 students), followed by biology (167 students), and medicine (148 students). Participants were enrolled in 20 different universities, which were all German universities, with the one exception of an Austrian university. A majority of 264 students studied at the same university. Between one and 49 students were enrolled at the other 19 universities. For a majority of 454 students the native language was German. Additionally, 27 students had German as well as another language as their native languages, 23 students had another language as their native language, and three participants did not answer this question. The students received 10€ for their participation.

3.2.2 Procedure

In order to reach the most possible participants we conducted the study both in a paper-and-pencil and an online version. Several studies showed that this should not lead to any major difficulties for our purpose (Bayazit & Aşkar, 2012; Preckel & Thiemann, 2003; Vleeschouwer et al., 2014). In the end, 237 students completed the paper-and-pencil version and 270 students completed the online version. The paper-and-pencil version was administered directly after the students attended one of their classes. The online version was sent out via various university mailing lists. The students could work for up to 25 minutes on the scientific reasoning test. The time limits for the figural, verbal, and numerical reasoning tests were 4.5, 3, and 5 minutes, respectively. Students had to complete the science knowledge test in 8 minutes. Finally, medical students could work on the error detection for

as long as they wanted. Adding the time for test instructions the total test time was about 60 minutes and about 10 minutes more for medical students. Data collection took place between October 2015 and March 2016.

3.2.3 Test instruments

For the assessment of scientific reasoning we used the updated version of the CTSR (Lawson, 2000), which is available online. This version of the test has been used in several studies, both with science majors and non-majors (Bao et al., 2009; Coletta & Phillips, 2005; Lawson, Clark, et al., 2000; Lawson, Alkhoury, et al., 2000). The questions were translated into German. Most of the 24 questions of the test come in pairs: First, students have to select a correct statement. Second, they have to select the correct justification for the statement. Questions have between three and five answer options. We scored the test as the author suggested: We combined the question pairs into one item with the exception of the last two questions, which are scored as separate items. This results in a maximum score of 13. The combination of two questions into one item reduces the guessing probability which makes it easier to construct a fitting Rasch model (Rost, 2004).

The context of the items was rated regarding its domain dependency by two members of the biology department and two members of the physics department. Three of the raters were postdoctoral researchers and one was an advanced PhD student. The rating scheme, which can be found in Appendix B, consisted of four categories: When an item had no physics or biology context it was given a 0 as a rating. When the item was embedded in a physics or biology context but the experts saw no domain-specific advantage it was rated with a 1. When the experts saw a domain-specific advantage that would help to solve the item but the item could also be solved with a potentially domain-general skill, like interpreting data, the item was given a 2 as a rating. Last, if it was absolutely necessary to master a

domain-specific aspect in order to solve the item it was rated with a 3. If the context of an item was mostly related to one subject, e.g. physics, but also had some relation to the other subject area, i.e. biology, only the main subject area was rated, which would be physics in this case. Table 4 contains the item numbers, the questions the items consist of, short item names, the descriptions of the items (readers can get a rough idea what is tested in the items based on these descriptions but it is not the exact wording of the questions), and the context ratings. As can be seen from the item descriptions, Items 1 and 2 strongly refer to physics concepts (conservation of weight and volume displacement), Items 3, 4, 8, and 9 rely strongly on quantitative reasoning and the remaining items cover core scientific reasoning skills such as generating and evaluating evidence as well as drawing conclusions.

In order to test general reasoning we took 10 items each from three subscales of a German intelligence test (IST 2000 R; Amthauer, Brocke, Liepmann, & Beauducel, 2001). The subscales tested figural, verbal, and numerical reasoning, respectively. The figural reasoning scale was about dice rotation, the verbal reasoning scale required the completion of a sentence in a logic way, and in the numerical reasoning scale test takers had to continue a row of numbers according to a hidden rule.

We tested science knowledge with 12 items from the National Assessment of Educational Progress for the 12th grade (U.S. Department of Education, Institute of Education Sciences, National Center for Education Statistics, National Assessment of Educational Progress [NAEP], 2000, 2005, 2009), which we translated into German. The 12 items were of either medium or hard difficulty and four items each were drawn from the three content areas physical science, life science, and earth and space science. In these items students had to, for instance, identify an ion with a negative charge, recognize a false statement about evolution theory, or state a universal feature of all galaxies. The knowledge required for answering these items was not addressed by any items in the Lawson test.

Table 4

Overview of Scientific Reasoning Test Items and Their Context Rating

Item nr.	Question	Item	Item description	Context rating[a]
1	1 + 2	Clay balls	What happens to the weight of a clay ball when the ball is flattened?	3 (Phy)
2	3 + 4	Marbles	When a steel marble is put into a cylinder filled with water, what happens to the water level compared to a glass marble?	3 (Phy)
3	5 + 6	Water tubes 1	How high will water rise in a narrow cylinder when water from a wide cylinder is poured into it?	0
4	7 +8	Water tubes 2	How high will water rise in a wide cylinder when water from a narrow cylinder is poured into it (the amount of water is different from the item before)?	0
5	9 + 10	Strings	Which strings (out of 3 possible strings) have to be used to find out if the length of a string has an effect on the time of one swing of the string?	1 (Phy)
6	11 +12	Flies 1	What does a graphic (showing the results of an experiment with flies) tell you about two possible influences on the behavior of flies?	1 (Bio)
7	13 + 14	Flies 2	This item is the same as item 6 with one influence factor and the results being changed.	1 (Bio)
8	15 + 16	Urn 1	What are the chances of drawing a certain kind of wooden piece out of a given set of pieces?	0
9	17 + 18	Urn 2	This item is the same as item 8 with the target piece and the set of pieces being changed.	0
10	19 + 20	Mice	Based on a figure about mice that carry a combination of two possible traits: Is there a link between the two traits?	2 (Bio)
11	21 + 22	Candle	How can a suggested explanation for a given observation be tested and which result would show that the explanation is wrong?	1 (Phy)
12	23	Blood cells 1	Which result of an experiment would show that the explanation for an observation is wrong?	1 (Bio)
13	24	Blood cells 2	Which result of an experiment would show that the explanation for an observation is wrong (the explanation is different from the item before)?	1 (Bio)

[a]Context rating: 0 – no item context from physics or biology; 1 – item has physics or biology context but no domain-specific aspects that are helpful for solving the item; 2 – domain-specific aspects are helpful in solving the item, but item can also be solved by a domain-general skill; 3 – mastery of domain-specific aspects are necessary for solving the item. Phy = physics; Bio = biology.

The error detection test for medical students was about finding errors in a diagnosis (Strobel, Heitzmann, Strijbos, Kollar, & Fischer, 2016). Students worked through two cases, in which they were presented with the data of patients and the subsequent diagnosis of a physician, which contained several mistakes. In total, students had to find 16 errors. For physics and biology graduate students we also collected data on the GPA of their undergraduate studies and their bachelor thesis grade in the form of a self-report by the students. This was not done for medical students since studying medicine in Germany is not organized in a two-tier educational system. A lower number in the GPA and the thesis variables represents a better grade in Germany. The complete test booklet can be requested from the author of this thesis.

3.2.4 Analysis

DIF was analyzed for the complete test (Strobl et al., 2015) as well as on the item level (Tutz & Berger, 2016; Tutz & Schauberger, 2015). Parameter estimation is difficult for items that were solved by every participant (or every participant in a subgroup in a DIF analysis) as well as for participants who solved all items (Rost, 2004). Thus, in all DIF analyses the first item was left out to avoid any problems caused by the fact that all physics students had correctly solved the item. Similarly, all students with a perfect score in the scientific reasoning test were excluded from the DIF analyses, which resulted in a sample of $n = 461$ for the DIF analyses. The three majors were represented in the DIF analyses with two dummy variables. We also evaluated the differences in scientific reasoning between majors with the help of item difficulties, an ANOVA, and an ANCOVA in which the general reasoning scales were the control variables. The ANOVA was followed up by post hoc tests with a Bonferroni correction.

To evaluate the independence of scientific reasoning from general reasoning within the bifactor model, which contained scientific reasoning and the three general reasoning scales, we calculated the Haberman criterion and omega hierarchical (omegaH) based on the instructions by Reise et al. (2013). The Haberman criterion is checked by comparing the proportional reduction in mean squared error (PRMSE) of the subscale ($PRMSE_S$) and the total score ($PRMSE_{TOT}$)[4]. If the $PRMSE_S$ (which is actually the same as Cronbach's alpha) is higher than $PRMSE_{TOT}$ the minimum criterion for interpreting a subscale score independently of a general factor is met. OmegaH represents the percentage of the variance in a test score that can be attributed solely to the general factor. It can also be calculated for a subscale score. In this case it represents the percentage of variance in a subscale score that can be attributed solely to the group factor (which is a subscale specific factor), and it is denoted as omegaS. The bifactor model itself was based on all 43 items from the scientific reasoning and the general reasoning scales. In the bifactor model all items load on two factors: the general factor and one of four group factors, and there is one group factor for each scale. All of the factors are being modelled as uncorrelated because the shared variance of the group factors is represented by the general factor. The fit of the bifactor model was assessed with the robust estimator provided by the R package lavaan (Rosseel, 2012).

In addition to the analyses based on the bifactor model, the three general reasoning scales predicted scientific reasoning in both a manifest and a latent regression. The latent regression was designed to reflect the manifest regression as close as possible. To calculate the latent regression, scientific reasoning, figural reasoning, verbal reasoning, and numerical reasoning were treated as latent variables, each connected with a single indicator, the sum

[4] It should be noted that not only the $PRMSE_S$ but also the $PRMSE_{TOT}$ is different for every subscale. As was stated in the Introduction, the comparison of the two values answers the question of whether observed subscale scores are better than the total score in predicting true subscale scores. Since the total score will predict the true subscale score to a different amount for every subscale, both values will be different for every subscale.

score of the according scale. The three latent general reasoning variables were correlated and individually predicted scientific reasoning. As suggested by Hayduk (1987) the residual variance of the scale scores got fixed to a value that was obtained by subtracting the scale reliability from 1 and multiplying this result with the scale variance. The omega reliability was used as the scale reliability for this calculation. Omega seemed better suited than Cronbach's alpha as the basis of this calculation for the latent regression – which used the sum scores of the reasoning scales – since omega expresses how precise a latent variable is represented by the sum score of a test (Revelle & Zinbarg, 2009).

The differing results regarding the factorial structure of both the CTSR and scientific reasoning tests in general led us to use an exploratory factor analysis (EFA) instead of a confirmatory factor analysis (CFA) to evaluate the factorial structure of the test. The number of extracted factors was determined by two common methods: the minimum average partial (MAP) test and a parallel analysis (Bühner, 2011). We used an EFA for dichotomous data with an oblique rotation using the oblimin rotation method. We applied an oblique rotation because we expected the factors to correlate (in case multiple factors would emerge).

We analyzed criterion validity with three regression analyses. The dependent variables in the three regressions were error detection, undergraduate GPA and the bachelor thesis grade, respectively. The five independent variables in each of these three regressions were scientific reasoning, the three general reasoning scales, and science knowledge. The regression with error detection was only conducted in medicine and the analyses with the undergraduate GPA and thesis grade were only conducted in physics and biology due to the constraints mentioned in the previous section about the test instruments.

Descriptive analyses including scale characteristics according to CTT, *t* tests (for the comparison of the paper-and-pencil and online versions), manifest regressions, the ANOVA, and the ANCOVA were calculated with IBM® SPSS® 23. The latent regression was

calculated in IBM® AMOS® 22. All DIF analyses, the EFA, and the bifactor model were calculated in R, version 3.2.5. We used the following packages: psychotree (Strobl et al., 2015), DIFtree (Berger, 2016), DIFlasso (Schauberger, 2016), Psych (Revelle, 2016), and Lavaan (Rosseel, 2012). If the Levene test was significant a correction was applied to the degrees of freedom for all *t*-test analyses (Bühner & Ziegler, 2009).

3.3 Results

3.3.1 Comparison of the paper-and-pencil test with the online version

We compared the results from the paper-and-pencil test and the online version for the main scales of this study (scientific reasoning, the three general reasoning scales, and science knowledge). The comparison was non-significant for scientific reasoning, verbal reasoning, and science knowledge. The highest *t* value of those comparisons occurred in the verbal reasoning scale, $t(487) = .06$, $p = .530$. A significant difference was found for the figural reasoning scale, $t(441.27) = 5.98$, $p < .001$, $d = 0.57$, and the numerical reasoning scale, $t(490) = 2.10$, $p = .036$, $d = 0.19$. However, the relation between these two scales and scientific reasoning was similar for the two test versions. Since determining this relation is what the general reasoning scales are mainly used for and because the overall differences are rather small we went on with the other analyses, combining the results from the two versions.

3.3.2 Descriptives

The means and standard deviations of the five main scales were as follows: Scientific reasoning had a mean of $M = 9.91$ ($SD = 2.22$), figural reasoning had a mean of $M = 6.29$ ($SD = 2.17$), verbal reasoning had a mean of $M = 7.54$ ($SD = 1.72$), numerical reasoning had a mean of $M = 7.76$ ($SD = 2.24$), and science knowledge had a mean of $M = 8.94$ ($SD = 1.94$).

The maximum total scores on the scales were 13, 10, 10, 10, and 12, respectively. Numerical reasoning had a heavily skewed distribution, *Mod* = 10. Lastly, error detection had a mean of *M* = 4.09 (*SD* = 2.00). The maximum total score was 16.

3.3.3 Domain generality analysis

The item discrimination and item easiness (which is 1 minus item difficulty) of the scientific reasoning items is displayed in Table 5. Item easiness is displayed for the whole sample and separately for each major. All items were easier for physics students compared to biology students.

Table 5

Item Discrimination and Item Easiness of the Scientific Reasoning Items

Item nr.	**Item**	**Item discrimination**	**Item easiness[a]**			
			Complete sample	Physics	Biology	Medicine
1	Clay balls	.12	.98	1.00	.95	.99
2	Marbles	.17	.88	.94	.85	.83
3	Water tubes 1	.32	.76	.87	.69	.70
4	Water tubes 2	.16	.53	.64	.43	.49
5	Strings	.19	.89	.93	.81	.92
6	Flies 1	.22	.44	.47	.46	.39
7	Flies 2	.33	.65	.71	.62	.59
8	Urn 1	.27	.90	.95	.83	.91
9	Urn 2	.35	.82	.89	.76	.79
10	Mice	.31	.70	.72	.62	.77
11	Candle	.43	.73	.86	.63	.66
12	Blood cells 1	.32	.78	.84	.67	.82
13	Blood cells 2	.24	.86	.86	.79	.93
	Mean	**.26**	**.76**	**.82**	**.70**	**.75**

[a]Item easiness = 1 – item difficulty.

We predicted scientific reasoning with the three general reasoning scales plus science knowledge in the overall sample and for each major separately. The regression weights for these regressions can be found in Table 6. The same pattern emerged in the overall sample and for the three majors: Science knowledge and verbal reasoning were stronger predictors than numerical reasoning and figural reasoning. Figural reasoning was a significant predictor in physics only (apart from the complete sample where all four scales had a significant regression weight), and numerical reasoning was a non-significant predictor in physics only.

Table 6

Regression Weights for Predicting Scientific Reasoning With General Reasoning Scales and Science Knowledge

Predictor	**Complete sample (*n* = 476)**		**Physics (*n* = 177)**		**Bio (*n* = 153)**		**Med (*n* = 146)**	
	β	*p*	β	*p*	β	*p*	β	*P*
Figural reasoning	.09	.017	.16	.022	.03	.703	.08	.317
Verbal reasoning	.27	< .001	.29	< .001	.27	< .001	.24	.002
Numerical reasoning	.17	< .001	.09	.177	.19	.010	.15	.046
Science knowledge	.30	< .001	.26	< .001	.26	.001	.30	< .001

An ANOVA comparing the three majors was significant with a medium to large effect size, $F(2, 503) = 24.82$, $p < .001$, $\eta^2 = .09$. Post hoc tests with a Bonferroni correction revealed that physics students were significantly better than medical students, who in turn were significantly better than biology students. The effect stayed significant after controlling for the three general reasoning scales in an ANCOVA, but the effect size was reduced to partial $\eta^2 = .05$.

In the tree based analysis for the whole test significant splits were found first for physics (vs. all remaining students) and within the non-physics branch for biology vs. medicine, which is the maximum number of splits for three groups. The results of the analysis can be seen in Figure 2. It shows the splits including *p* values and item difficulty profiles. This tree based analysis only tested if there were differences overall, i.e. differences between single items should be interpreted with caution because no tests of significance were calculated on the item level. On the overall level the results of the analysis mean that the test is producing biased results for comparisons of any two of the three majors.

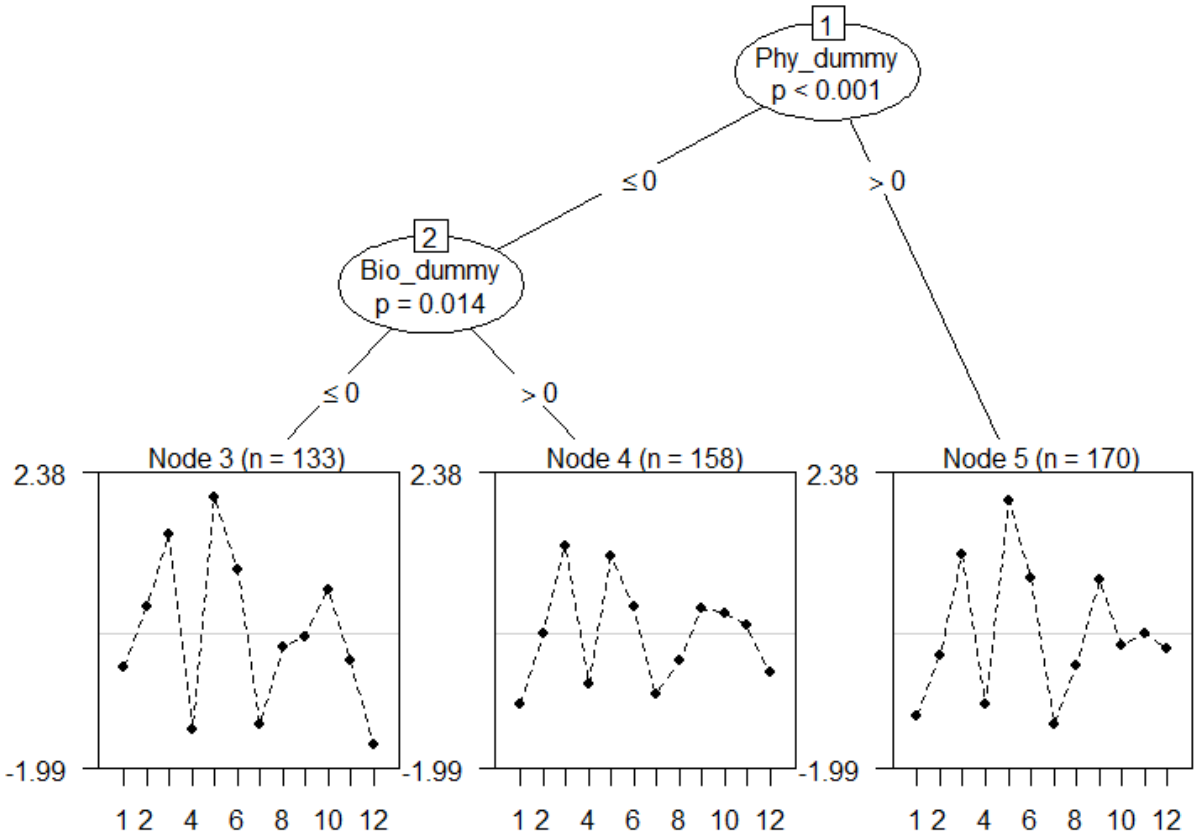

Figure 2. Result of the tree based analysis for the whole scientific reasoning test. Since Item 1 was excluded the number 1 on the x-axis relates to Item 2, number 2 relates to Item 3 and so on. Node 3 represents medicine, node 4 represents biology, and node 5 represents physics.

To see which items were responsible for this breach of measurement invariance we needed the results of the item-focused tree and lasso based analyses. Table 7 shows the item difficulties resulting from these two analyses. When interpreting these values one should remember that the absolute differences between the majors are underestimated in the lasso analysis. The direction of the bias is of higher importance in the lasso based analysis.

Table 7

Item Difficulty Parameter From the Tree and Lasso Based Analyses

Item nr.	**Item**	**Tree based analysis**			**Lasso based analysis**			**Context rating**
		$Diff_S$	$Diff_{Phy}$	$Diff_{Bio}$	$Diff_S$	$Diff_{Phy}$	$Diff_{Bio}$	
1	Clay balls	na	na	na	na	na	na	3 (Phy)
2	Marbles	-1.15			-0.59			3 (Phy)
3	Water tubes 1	-0.13			0.42			0
4	Water tubes 2	1.29			1.83			0
5	Strings	-1.27			-0.70			1 (Phy)
6	Flies 1	2.18		1.03	2.31	2.41	2.07	1 (Bio)
7	Flies 2	0.87		0.18	1.17	1.19	1.10	1 (Bio)
8	Urn 1	-1.39			-0.82			0
9	Urn 2	-0.56			0.00			0
10	Mice	-0.04	0.88		0.84	1.02	0.87	2 (Bio)
11	Candle	0.12			0.67	0.62	0.67	1 (Phy)
12	Blood cells 1	-0.23			0.32			1 (Bio)
13	Blood cells 2	-1.99	-0.23	-0.89	-0.36	-0.22	-0.30	1 (Bio)

Note. $Diff_S$ = baseline difficulty of the sample; $Diff_{Phy}$ = difficulty for physics students; $Diff_{Bio}$ = difficulty for biology students. If a cell is empty the difficulty for that group is the same as for the whole sample. In case the difficulty diverges for both physics and biology students, $Diff_S$ becomes the difficulty for medical students. Item 1 was left out of the analyses because 100% of physics students solved the item. Lower numbers indicate a lower difficulty.

Both methods agreed that Items 6, 7, 10, and 13 were biased. The lasso based analysis also indicated a bias in Item 11. When we compared physics and biology students in terms of which group has an advantage, we saw that the bias always corresponded with the context judgment. Biology students had an advantage in Items 6, 7, 10, and 13. Physics students had

an advantage in Item 11. Items 2 and 5, which also had a subject specific context, showed no DIF but were also rather easy items. Of the items included in the DIF analysis, only Item 8 was easier than these two items. Item 12 was the only item that showed DIF despite both having a subject specific context and not being among the easiest items. Item 12 had a medium difficulty relative to the other items.

3.3.4 The connection of scientific reasoning with general reasoning and science knowledge

The first relevant result for judging the relation between scientific reasoning and general reasoning as well as science knowledge is the regression, in which scientific reasoning was predicted by the other main scales. We saw the regression weights of this regression in Table 6. The explained variance for only the general reasoning scales was $R^2 = .21$. Together with science knowledge this value rose to $R^2 = .30$. Science knowledge had the single highest regression weight. In the latent regression the amount of variance explained by the three reasoning scales rose to $R^2 = .36$. The latent regression weights for figural reasoning, verbal reasoning, and numerical reasoning were as follows: $\beta = .13$, $p = .013$ (figural reasoning), $\beta = .46$, $p < .001$ (verbal reasoning), and $\beta = .23$, $p < .001$ (numerical reasoning).

The Haberman criterion was met by the scientific reasoning scale, since $PRMSE_S = .63 > PRSME_{TOT} = .56$. OmegaS for scientific reasoning (the percentage of variance in the scientific reasoning score that can be attributed solely to scientific reasoning and not general reasoning) was .002, so less than 1%. For the other three reasoning scales the values were 71.9% (figural reasoning), 30.2% (verbal reasoning), and 72.0% (numerical reasoning).

The factor loadings and the uniqueness of the scientific reasoning items that resulted from the bifactor model can be seen in Table 8. In the bifactor model, which separated a

general factor from the four group factors (scientific reasoning and the three general reasoning scales), the majority of scientific reasoning items had a negative loading on the scientific reasoning subscale. The items that kept a positive loading on the scientific reasoning subscale after the separation of the general factor are Items 5, 10, 11, 12, and 13. Items 12 and 13 are the only two items that had a higher loading on scientific reasoning than on the general factor. Most items also had a high uniqueness value, so a high percentage of their variance that was explained by neither the general factor nor the scientific reasoning factor. The fit of the bifactor model, which was based on a sample of $n = 479$ students, was as follows: $\chi^2(817) = 956.29$, $p = .001$, CFI = .94, RMSEA = .02, and SRMR = .093.

Table 8

Factor Loadings and Uniqueness of Scientific Reasoning Items Based on the Bifactor Model

Item nr.	**Item**	**Factor loading**		**Uniqueness**[a]
		General factor	Scientific reasoning	
1	Clay balls	.47	-.21	.73
2	Marbles	.38	-.07	.85
3	Water tubes 1	.59	-.23	.60
4	Water tubes 2	.35	-.13	.86
5	Strings	.46	.10	.78
6	Flies 1	.24	-.14	.92
7	Flies 2	.44	-.12	.79
8	Urn 1	.61	-.19	.60
9	Urn 2	.54	-.23	.65
10	Mice	.49	.09	.75
11	Candle	.71	.08	.48
12	Blood cells 1	.49	.58	.42
13	Blood cells 2	.46	.78	.18

[a]Uniqueness is the percentage of variance of an item that is explained by neither factor.

3.3.5 The factorial structure of scientific reasoning

All analyses in this section were calculated with $n = 506$. Using Cronbach's alpha the reliability of the scientific reasoning scale was .63. The item discrimination of all items can be seen in Table 5. The average item discrimination was .26. The average inter-item correlation was .11 with a range from -.02 (Item 1 by Item 10) to .44 (Item 12 by Item 13). Two correlations between items were negative (Item 1 by Item 6 and Item 1 by Item 10) and the second highest correlation was .29 (Item 8 by Item 9).

The most relevant analysis regarding the factorial structure of scientific reasoning is the EFA. Here, the MAP test suggested a one-factor model, but the PA suggested a seven-factor model. Both models were subsequently calculated. The variance explained by the one-factor model was 24%. The loadings ranged between .71 (Item 11) and .27 (Item 4). The variance explained by seven factors was 66%. The factor loadings of the seven-factor solution are included in Table 9. When items were assigned to factors based on their highest loading, six factors consisted of two items and one factor consisted of just one item (Item 11). Factors 1, 3, and 4 combined items that share the same basic task (blood cell experiments, water tubes, and experiments with flies).

3.3.6 Criterion validity

In physics the relation between scientific reasoning and undergraduate GPA was non-significant, $\beta = .21$, $p = .101$, $n = 82$. In biology the relation was significant $\beta = -.36$, $p = .014$, $n = 65$. The same picture emerged for the relation with bachelor thesis grades. In physics it was non-significant, $\beta = .08$, $p = .529$, $n = 84$, but in biology it was significant $\beta = -.30$, $p = .034$, $n = 75$.

When predicting error detection with the five main scales in a sample of n = 122 medical students, no scale had a significant regression weight, R^2 = .05, p = .364. The regression weight of scientific reasoning was close to zero, β = -.002, p = .983.

Table 9

Factor Loadings for the Seven-Factor Model Based on the Exploratory Factor Analysis

Item nr.	**Item**	**Factor loadings on respective factors**						
		1	2	3	4	5	6	7
12	Blood cells 1	**.93**	-.15	.04	.10	-.05	.08	.18
13	Blood cells 2	**.82**	.24	-.04	-.11	.10	-.01	-.26
1	Clay balls	.01	**.96**	.04	.11	-.05	-.01	-.01
8	Urn 1	-.02	**.38**	.16	.08	**.32**	.11	.24
3	Water tubes 1	-.01	.03	**.98**	-.05	-.05	.06	-.01
4	Water tubes 2	.09	.13	**.32**	.07	.10	-.20	.15
7	Flies 2	.01	.10	-.06	**.98**	-.02	.04	.00
6	Flies 1	-.05	-.17	.08	**.37**	.11	.12	-.05
5	Strings	.00	.03	-.20	-.10	**.76**	.19	.07
10	Mice	.10	-.23	.24	.22	**.62**	-.14	-.16
11	Candle	.15	-.04	.18	.15	.08	**.65**	-.07
2	Marbles	-.09	.07	.16	.10	.08	.24	**-.47**
9	Urn 2	.05	**.32**	.24	.04	.28	.12	**.40**

Note. Loadings above a value of .3 are marked in bold.

3.4 Discussion

This study evaluated the domain generality assumption, the divergent validity, the factorial structure, and the criterion validity of a commonly used scientific reasoning test. The scientific reasoning test, along with three general reasoning scales and a science knowledge test, was administered to 507 university students who studied physics, biology, or medicine.

The IRT based analyses of the domain generality assumption revealed that there were four to five biased items. The biased items were mostly items with a biology context that

gave biology students an advantage over physics students. This bias was not found in the analyses based on CTT.

In a regression analysis, science knowledge and verbal reasoning were stronger predictors of scientific reasoning compared to numerical reasoning and figural reasoning. General reasoning explained 21% of the variance in scientific reasoning in a manifest regression and 36% in a latent regression. In a bifactor model, in which scientific reasoning and the three other reasoning scales were all treated as subscales of general reasoning, scientific reasoning did not stand out as particularly independent from general reasoning. While the scientific reasoning scale met the Haberman criterion, less than 1% of its variance was independent of general reasoning, and after controlling for the general factor several items loaded negatively on the scientific reasoning scale.

An EFA suggested either a one-factor or a seven-factor model as the best fitting models. In the seven-factor model several factors were determined by a shared task setting. Several other results supported the notion that there is not a strong homogeneity between scientific reasoning items.

The criterion validity varied between the majors. While no significant relations were found for either physics or medicine, biology students' undergraduate GPA and bachelor thesis grade were significantly predicted by scientific reasoning. The following four sections will analyze the main results in detail and discuss implications. This will be followed by sections that address the limitations, suggestions for further research, and, last, the main conclusions of the study.

3.4.1 Domain generality vs. domain specificity of scientific reasoning

Before we analyze the results regarding the domain generality of the test in detail it should be noted that the amount of biased items was rather under- than overestimated since

the IRT methods that were used are known to produce few false alarms, but sometimes too few hits, in samples of a size comparable to this study (Tutz & Berger, 2016; Tutz & Schauberger, 2015). Anybody, who wants to draw conclusions from the presented results, should keep that in mind.

At first glance it looks like there is no straight forward relation between the bias that was found and the context rating that was based on the model by Shavelson and Huang (2003): While the items with DIF indeed displayed a bias that corresponded to their context rating, there were also items, e.g. Item 2, that had an even stronger domain-specific influence according to the expert rating but did not show any bias. However, this can possibly be explained by the fact that these items without DIF had a low difficulty. It is possible that all participants had already mastered the domain-specific aspects that are necessary for solving the items. For instance, Item 2 can only be solved with domain-specific knowledge about volume displacement. This physical knowledge might be so basic, though, that no measurement invariance occurred. DIF might occur only in items of a higher difficulty. The sole exception that is not captured by this explanation is Item 12.

Another aspect that is aligned with the model of Shavelson and Huang (2003) is that no DIF was found for items relying heavily on quantitative reasoning. As the model would predict, quantitative reasoning is on a higher and broader level, which does not vary between domains. In contrast, the items that did display DIF were items that measured core aspects of scientific reasoning such as generating and evaluating evidence or drawing conclusions.

It should be pointed out that the bias was not apparent in any way when we just looked at the item difficulties. More strikingly, comparing the means of groups, a method that several authors used to evaluate their domain generality and domain specificity claims in the past (Cloonan & Hutchinson, 2011; Weld et al., 2011), would have even led to the opposite conclusion of the DIF analysis, namely that physics students had an advantage. However,

when we take the IRT analyses into account it rather seems like physics students have higher latent scientific reasoning skills and the difference in the mean test score would be even bigger if the items were not biased in favor of biology and medical students. Additionally, we did not see a difference in the pattern of the regression results for the different majors when we predicted scientific reasoning with the general reasoning scales and science knowledge. However, it is not very surprising that we did not find measurement invariance, but predictive invariance, which is about group differences in the relation between tests, still seemed to hold: Millsap (Millsap, 1995) showed that measurement invariance and predictive invariance cannot be achieved at the same time under realistic circumstances. Furthermore, it has been argued that measurement invariance should be preferred over predictive invariance (Molenaar & Borsboom, 2013).

Taken together, what are the consequences of the results from the domain generality evaluation? In the answer to this question we have to consider that the CTSR, which is based on the theory about the stages of the development of human thinking (Inhelder & Piaget, 1958), is a scientific reasoning test that is conceptually rather close to general reasoning. This study showed that there is also a sizable empirical overlap with general reasoning. The influence of domain-specific knowledge on the test is considered to be “minimal” (Osborne, 2013, p. 269). When we find DIF between domains in such a test it demonstrates that it will be hard to measure scientific reasoning without any domain-specific influence. This also indicates that scientific reasoning skills are probably not a completely domain-general construct. In terms of test evaluation the results show the benefits of using IRT based methods on top of CTT based methods. Otherwise we could miss crucial information about the measurement invariance of scientific reasoning tests.

When we consider the practical implications for using scientific reasoning tests we should remember that the presence of bias does not automatically mean that the bias will be a

concern in the application of the test (Borsboom, 2006). The differences between the majors did not disappear, so the advantage of biology and medical students was not enough to compensate for or reverse their lower latent skill level. In cases in which we are mainly concerned about the relationship with other variables, scientific reasoning tests might still produce unbiased predictions.

However, it also became very clear that the item context has to be considered when comparing university students with different majors. Even if there are no apparent domain-specific elements required for solving the item, just embedding an item into a specific context domain can be enough to induce bias. The largest difference in item difficulty was more than 1.6 logits between medicine and physics in Item 13 which is more than half of the range of the difficulties of items without DIF. This is not something that can be easily neglected. High-stake decisions in which students with different majors are compared (e.g. selecting applicants for an interdisciplinary PhD program) and which are based on the CTSR or comparable tests would be at least doubtful. This is especially true when domains would be compared that are further apart then the rather closely connected domains of physics and biology.

We should be careful though to simply scratch all biased items from scientific reasoning tests. These items might still cover important aspects of scientific reasoning. They should be included in tests, but carefully interpreted. An alternative solution, which is especially interesting for tests with a large number of items, would be to use the remaining unbiased items as anchors for estimating the difficulty of the biased items separately for every domain (Boone et al., 2014).

3.4.2 The connection of scientific reasoning with general reasoning and science knowledge

An overall evaluation of the connection between scientific reasoning and general reasoning gets complicated by the differing results. On the one hand, the manifest and latent regressions imply that despite a significant overlap between the two constructs, at least around two thirds of the variance of scientific reasoning is still left unexplained. The size of this overlap is remarkably similar to prior studies that evaluated the relation between general reasoning and scientific reasoning in younger populations, both for the manifest and the latent regression (Klos et al., 2008; Koerber et al., 2015; D. Mayer et al., 2014). On the other hand, the omegaH calculations show that there is hardly any aspect of scientific reasoning that goes beyond general reasoning. The key to resolving this difference might be found on the item level of the bifactor analysis. Items that we view as core scientific reasoning items mostly keep their positive loadings on the scientific reasoning factor. This is not true for the items that rely heavily on quantitative reasoning. If this result holds up in further research it would be an argument against including quantitative reasoning into scientific reasoning conceptualizations. Additionally, one other aspect is noteworthy regarding the connection between scientific reasoning and general reasoning: The high importance of verbal reasoning is a reminder that language requirements should be either kept at a minimum in scientific reasoning tests or that verbal reasoning has to be actively controlled.

In line with prior studies about the CTSR we also found a significant relation between scientific reasoning and science knowledge (Lawson, Clark, et al., 2000; Lawson, Alkhoury, et al., 2000). The source for the relation, that was stronger than for any single general reasoning scale, is still unclear, especially considering that the knowledge in the science knowledge test was not the same as any knowledge that items of the CTSR might utilize. It seems like the students who possessed the necessary knowledge for the science knowledge

test also possessed the knowledge and the skills that are needed to perform well in the scientific reasoning test. Two explanations seem reasonable for this phenomenon: It is possible that higher scientific reasoning skills helped the students acquire science knowledge. Another explanation would be that interest in science is a hidden mediator. This would make sense, considering that more knowledge and interest in one domain goes along with more knowledge and interest in another domain (Alexander, Jetton, & Kulikowich, 1995).

3.4.3 The factorial structure of scientific reasoning

The EFA suggested either a one-factor model or a seven-factor model. Additionally the item uniqueness in the bifactor model was high and the inter-item correlation, the internal consistency, and the item discrimination were low. If we combine all these different results that touch on the factorial structure of scientific reasoning the picture emerges that there is a weak general scientific reasoning factor but also several minor, more specific factors. It seems clear that scientific reasoning is not just a simple, unidimensional construct. A connection between the different subskills does exist but it is a complicated one. The interpretation of the process that led to the grouping of the items is also not completely straightforward. Three of the factors appear to be organized by a shared or similar item prompt (blood cell experiments, water tubes, and experiments with flies). Furthermore, it is possible that difficulty factors, on which items get grouped together just because they have a similar level of difficulty, played a role here (McDonald, 1965). What we did not observe was a grouping by context or skill group: For instance, not all items with a biology context grouped together and neither did all items that rely heavily on quantitative reasoning group together. We also did not find a simple divide by different knowledge types as Li et al. (2006) suggested. If such a divide would have been the driving force behind our results we probably would have seen a clear split between the first two items and the rest of the items.

Of course it would be possible to dismiss the results as test or sample specific. However, as we saw in the Introduction, varying results regarding the factorial structure are rather common for scientific reasoning. The combination of a weak general scientific reasoning factor and several item-specific factors might be the explanation for this variety of results. Some analyses might have discovered the general factor while others found the item-specific factors. Such a structure would not be unique to scientific reasoning. It is not uncommon to find a similar structure in knowledge tests, such as the Force Concept Inventory (FCI; Hestenes, Wells, & Swackhamer, 1992). One EFA of the FCI revealed not one but rather nine to 11 factors (Huffman & Heller, 1995). The items of the FCI were only loosely connected. Similar to our results, questions with a similar prompt, e.g. items about a hockey puck or a rocket in space, formed their own exclusive factor instead of joining a broader factor, which would include all items that address the same underlying physical principle. In another study about the FCI a five-factor model was optimal but all factor models from a one-factor model to a 10-factor model would have fit the data (Scott & Schumayer, 2015).

What are the implications of these results? First of all, it seems important to always explore the factorial structure of scientific reasoning tests since it is likely that it will not be straightforward. Apart from learning about the factorial structure such an analysis is also helpful for finding potentially odd items. For instance, Item 4 showed a high uniqueness in both the EFA and the bifactor model and a closer inspection of the item suggested that one of the answer options might have been confusing. Similarly, the particularly high correlation between Items 12 and 13 provides an argument for scoring these two questions as a single item in the same way that the other 22 questions are combined, too.

Another implication is that a test that claims to measure scientific reasoning in a general way should try to include as many diverse aspects of the construct as possible. If this

is not done the test should state that it is not measuring scientific reasoning in general but rather a subset of scientific reasoning skills. A different approach would be to construct many tests, each measuring only one very narrow skill. However, such a test would always have the problem that skill specific variance would be confounded with the general scientific reasoning factor (Gustafsson & Åberg-Bengtsson, 2010). In order to get a good estimation for just the skill specific variance we would need to measure many scientific reasoning skills to control the influence that the general scientific reasoning factor has in each of these specific factors. In order to find out more about which specific factors exist, it might be a good idea to follow the approach that Adams and Wieman (2015) took in the field of problem-solving. They analyzed the problem-solving process in great detail and list a total of 44 subskills. Since very specific elements seem to play an important role in scientific reasoning tasks, too, such a detailed analysis of scientific reasoning tasks might be worthwhile.

3.4.4 Criterion validity

Unfortunately, the reduced sample size for all criterion validity analyses does not allow us to draw strong conclusions. Additionally, the results from the error detection test might be restricted by the fact that it was conducted at the very end of the testing procedure and test takers were not performing at their best level. This would explain the low average: Only about one quarter of the errors in the diagnoses were found on average. The only tentative conclusion we can draw is that the connection between scientific reasoning and external variables is stronger for biology than for physics students. More studies are needed to confirm this, though, before an accurate interpretation is possible.

We also need to know more about the predictive validity of scientific reasoning tests. It would be especially interesting to observe if scientific reasoning skills support the acquisition

of content knowledge. One study that found a relation between CTSR scores and knowledge gains in physics concepts shows that such an endeavor is promising (Diff & Tache, 2007).

3.4.5 Limitations

One possible limitation is the high mean score that the students achieved in the scientific reasoning test. This result is not unique to our study. In another study, university students also correctly solved around three quarters of the CTSR items (Bao et al., 2009). Thus, it seems like we need more difficult items for testing scientific reasoning at the university level.

Compared to another study about the CTSR, which showed an internal consistency of .81, the reliability in our sample was considerably lower (Lawson, Alkhoury, et al., 2000). However, the reliability value of .63 that we found is also not unheard of for the CTSR. In a sample of 667 non-science majors the test-retest reliability was .65 (Lawson, Clark, et al., 2000).

Another problem could be the two versions of the test. However, the results obtained from the online and the paper-and-pencil versions were mostly comparable. The fact that the differences for three of five scales were non-significant in samples of around 500 students, in which small effects can be discovered, demonstrates a remarkable similarity. The problem with figural reasoning, which showed the highest difference between the two versions, probably was that the prompt for the task (a selection of dice that had to be compared with a different die in every item) had to be repeatedly displayed in the online version to avoid scrolling. Probably, this made the task easier.

Other limitations include the sample and the test selection. It is possible that results would be different with students at the beginning of their studies when they are less emerged in one domain. Obviously, our results are limited to the CTSR to a certain degree. However,

considering how commonly used the CTSR is and that its items cover skills such as generating and evaluating evidence as well as drawing conclusions, which are commonly tested in scientific reasoning tests (Opitz et al., 2015), we are optimistic that our results are not irrelevant for the broader field of scientific reasoning assessment.

3.4.6 Suggestions for further research

One avenue for further research is the analysis of aspects that influence item difficulty apart from the construct that was intended to be measured. For instance, in their study about the FCI, Scott and Schumayer (2015) found one factor that was not determined by a target construct but rather by the fact that the items that loaded on the factor all required some kind of visual problem-solving. Another study highlighted the effects of specialist terms, answer option length, the work with tabulated data, and several other aspects on item difficulty (Stiller et al., 2016). Similar analyses are necessary for more scientific reasoning items considering the importance of item-specific factors. We have to find out if these items specific factors are genuinely interesting aspects of scientific reasoning or rather artefacts of superficial item characteristics.

This study demonstrated how the application of modern statistical methods in a detailed analysis of one test can help us learn lessons about the construct of scientific reasoning itself. This approach should be followed up in the future, especially since research about the assessment of higher education competencies has not made extensive use of complex analysis designs yet (Zlatkin-Troitschanskaia et al., 2015). This study has applied some modern analysis approaches but of course many other approaches might also be useful. Analyzing DIF within an IRT framework is also possible via Bayesian methods (Soares, Goncalves, & Gamerman, 2009). Besides, DIF can be analyzed within a factor analysis framework and recent methods like moderated nonlinear factor analysis try to overcome deficits of older

methods (D. J. Bauer, 2016). It would be interesting to compare the results from even more DIF analyses to see how stable the results are across methods. If the results from different analyses are stable we can have more confidence in the conclusions about the domain generality of scientific reasoning that we draw from DIF analyses.

The results from this study strongly indicate that scientific reasoning does not have a simple unidimensional structure. Thus, when we want to learn more about scientific reasoning by analyzing its factorial structure we need to use models that allow the integration of the complexities of the scientific reasoning construct. For instance, we could test predictions about item clusters with exploratory structural equation modeling, which seems to be superior for complex measurement models (Prudon, 2016). Once we know more about the structure, multidimensional IRT models would probably be more accurate than standard unidimensional models (Hartig & Höhler, 2009). Neumann, Neumann, and Nehm (2011) used such a multidimensional model in the analysis of a nature of science test, so it might be interesting for scientific reasoning as well. Both multidimensional IRT models and bifactor models can be used for longer tests with several items per assumed subskill to check if there is indeed a general scientific reasoning factor and not just several minor factors (Blömeke et al., 2015).

In terms of test construction and administration, multidimensional adaptive testing (MAT) could be a useful approach for the complex structure of scientific reasoning (Frey & Seitz, 2009). When a test is constructed based on the MAT approach test takers have to answer even less items compared to a standard adaptive testing setting to reach the same level of accuracy: Not only is the difficulty of the presented items adapted to the level of the test takers but the approach also recognizes the multidimensionality of the tested construct. However, using MAT requires that the dimensions and the covariance structure of the construct is known, something that cannot be said of scientific reasoning right now. So before

we could use the MAT approach we would have to find out more about the dimensionality of scientific reasoning and to what extend the different dimensions are connected.

3.4.7 Conclusions

To summarize, this study demonstrates the usefulness of combining different test analysis methods in order to learn as much as possible about scientific reasoning and its assessment. It also suggests new methods to test assumptions – e.g. the assumption of domain generality – that are superior to current standard methods. Besides, the study shows how much can be learned from, and how much still has to be learned about, current scientific reasoning tests and we recommend that test authors do this before constructing new scientific reasoning tests. For instance, we learned which items are at the core of scientific reasoning and which items are biased and this knowledge will be useful to improve future tests.

The study also contributes to the goal of understanding the construct of scientific reasoning better. One important lesson is that quantitative reasoning should rather be excluded from the conceptualization of scientific reasoning. Furthermore, we found more evidence that scientific reasoning is not a simple unidimensional construct, it is probably impossible to measure in a completely domain-general way, and it contains some aspects that are distinct from general reasoning but there is also a considerable overlap that definitely needs more attention. Overall, this study contributes to a better understanding of the assessment of higher education competencies and it presents ways to tackle the big research gap in the field that was noted by Zlatkin-Troitschanskaia et al. (2015).

4 General Discussion

The last part of this thesis is an overall discussion of the presented studies. The first section is a summary of the studies. The General Discussion continues with highlighting aspects from the presented studies that are relevant for the conceptualization of scientific reasoning and for evaluating and constructing scientific reasoning tests (theoretical and methodological implications) as well as aspects that are relevant for the use of scientific reasoning tests in practice (practical implications). This is followed by a summary of the limitations of the presented studies. Finally, this General Discussion is complemented by suggestions for further research.

The two sections about the implications of the results will provide answers to the questions raised in the Introduction from both of the two main goals of this thesis. The first goal was to explore the current state of scientific reasoning assessment and scientific reasoning conceptualizations. The second goal was to demonstrate how an in-depth analysis of an already existing scientific reasoning test can help us learn more about the scientific reasoning construct and what has to be improved in the construction of future scientific reasoning tests. The part about theoretical and methodological implications will address the questions from the first goal about how scientific reasoning is conceptualized, touching on the issues of which skills belong to scientific reasoning, how are the skills connected, and are they domain-general or domain-specific. The part also addresses the psychometric properties of scientific reasoning tests and the changes that occurred over time. Additionally, the part addresses the according questions from the second goal about domain generality, construct

validity, and criterion validity. Furthermore, the part reflects on the usefulness of the statistical methods employed in the second study and what we have learned from the presented studies for test construction. The topic of criterion validity also gets picked up again in the section about suggestions for future research. The section about practical implications touches on the question of which tests currently exist. It expands on this topic by stating the implications that the presented studies have on the question of which decisions should or should not be based on SR tests.

4.1 Summaries of the Presented Studies

4.1.1 Study 1

Study 1 reviewed the field of scientific reasoning tests. Its first goal was to provide an overview of existing scientific reasoning tests, their test formats, their target groups, and how test authors evaluated the psychometric properties of the tests. The second goal was to use the tests as proxies in order to learn more about the conceptualization of scientific reasoning. The focus was on the skills included in the tests, the theories about scientific reasoning underlying the tests, and which position the test authors took on the domain generality vs. domain specificity debate. The skills included in the test were sorted according to the scientific reasoning framework by F. Fischer et al. (2014), which recognizes eight scientific reasoning skills (*problem identification*, *questioning*, *hypothesis generation*, *construction and redesign of artefacts*, *evidence generation*, *evidence evaluation*, *drawing conclusions*, and *communicating and scrutinizing*).

The review found 38 tests, which are part of two waves of test development. The first wave was in the 1970s and 1980s and the second wave started with the beginning of the 21st century and is still ongoing. Most tests targeted students attending primary, secondary, or

tertiary education institutions. The review discovered that evaluations of the psychometric properties of the tests were lacking in relevant ways. Reporting reliability was not a standard. The evaluation of construct validity (including factorial, convergent, and divergent validity) and criterion validity had remarkable gaps: Few test authors explored the factorial structure of their tests, compared their results with other scientific reasoning tests, or tried to show differences to measures of other cognitive abilities. The attempt to determine the current state of the criterion validity of scientific reasoning tests was hindered by the absence of both common measures of criterion validity and of results about predictive validity. The variety of test formats is growing. While the first wave of tests mainly deployed multiple-choice tests, this is no longer true for the second wave.

Regarding the second goal, the review showed that *evidence generation* was the scientific reasoning skill that was tested most often. Other skills tested in more than half of the tests were *hypothesis generation*, *evidence evaluation*, and *drawing conclusions*. The overall number of tested skills had not increased in newer tests but with *formulating questions* one more skill was added to the rotation of tested skills. *Quantitative reasoning* was the most tested skill that is not included in the framework by F. Fischer et al. (2014). The theories used for test construction were either assuming a single general scientific reasoning dimension, several independent scientific reasoning skills, or several skills organized in a problem-solving process. Furthermore, many scientific reasoning conceptualizations used in the tests were inspired by educational guidelines. One third of the test authors assumed scientific reasoning to be domain-general, one third assumed it to be domain-specific, and the remaining third made no or no clear assumptions. Just one test made a more complicated assumption, which differed for different parts of the test. There was a shift in the tests from the second wave towards the assumption of domain specificity.

Overall, Study 1 showed that there is a sizeable selection of scientific reasoning tests one can choose from, which has widened in part due to a rising interest in scientific reasoning in recent years. Study 1 recommended that people in need of a scientific reasoning test should define requirements for a test before searching for a test. The shift towards more nuanced conceptualizations of scientific reasoning – which was reflected in the use of theories that recognize multiple skills, the growing popularity of the domain specificity assumption, and the higher number of test formats – was not reflected in the total number of skills that were tested. Furthermore, tests need to check their assumptions about their structure (including assumptions about domain generality or domain specificity) and use more advanced statistical methods in doing so. This is also true for the evaluation of criterion validity. Instead of developing more tests, Study 1 recommended to first learn more about the existing ones. Finally, Study 1 concluded that, until we know more about scientific reasoning and its assessment, we should refrain from making high-stake decisions based on scientific reasoning tests.

4.1.2 Study 2

Based on the conclusions from Study 1, Study 2 evaluated one scientific reasoning test in depth in order to learn more about scientific reasoning as a construct and to gather recommendations for future test evaluation and development. The main focus of Study 2 was on the domain generality vs. domain specificity debate, the divergent validity of the test, and its factorial structure. To test the domain generality assumption of the test, Study 2 presented the idea that the assumption can be phrased as a measurement invariance problem (at least at the university level) and deployed newly developed methods to check measurement invariance. Study 2 also evaluated the criterion validity of the test.

To reach these objectives the test was administered to 507 university students. University students were chosen because there is a call in the literature to focus more on the assessment of cognitive skills that are associated with a tertiary education. The participants studied medicine, physics, or biology. The Classroom Test of Scientific Reasoning (CTSR), also known as the Lawson test, was chosen to test scientific reasoning as it is a commonly used scientific reasoning test (Lawson, 1978, 2000). Additionally, the test contains several items with contexts from physics and biology. Students were also tested with three general reasoning scales – figural reasoning, verbal reasoning, and numerical reasoning – a science knowledge test, and – in the case of medical students – an error detection test. The error detection test as well as the bachelor thesis grades and the grade point average (GPA) of the undergraduate studies were used to establish criterion validity.

Analyses on the item level revealed that several items showed differential item functioning (DIF). In all of these items the influence of the context was as expected, e.g. biology students having an advantage over physics students in items with a biology context. In most cases, the biased items were just set in a domain context and there was no further influence of domain-specific content or skills, as judged by experts from the according fields. DIF was not present in items associated with quantitative reasoning. The bias was not discovered by methods from classical test theory (CTT): Although physics students were disadvantaged by the biased items, an analysis of the group means showed that they were still better than medical students, who in turn were better than biology students. Furthermore, all item difficulties, which were low overall, were even lower for physics students compared to biology students. The relation between scientific reasoning and the general reasoning scales was comparable for the three majors.

In terms of divergent validity, a manifest regression showed that the three general reasoning scales explained 21% of the variance in scientific reasoning. In a latent regression

this value rose to 36%. When scientific reasoning and the three general reasoning scales were treated as four subscales of general reasoning within a bifactor model, scientific reasoning passed the Haberman criterion, a minimum criterion that a test should meet in order to be interpreted as a scale that has merit beyond a more general factor. However, in the bifactor model hardly any variance of the scientific reasoning items could be attributed to a scientific reasoning factor once the general reasoning factor of the model gets taken into account. Most items that could be described as representing core scientific reasoning skills according to F. Fischer et al. (2014) retained a positive connection with the scientific reasoning factor after controlling for general reasoning. Science knowledge had a stronger connection with scientific reasoning than any of the three general reasoning scales.

Regarding the factorial structure of scientific reasoning, an exploratory factor analysis suggested a solution with either one factor – according to the minimum average partial (MAP) test – or with seven factors – according to the parallel analysis. In the seven-factor solution, items were mainly grouped by shared task scenarios or by item difficulty. The one-factor solution explained 24% of the variance and the seven-factor solution explained 66%. The results from the factor analysis were complemented by low item discrimination, low reliability, and low correlations between the items. In terms of criterion validity, significant results were only found for biology students. Within their subsample, higher scientific reasoning skills correlated with a better GPA and a better bachelor thesis grade.

Study 2 concluded that a completely domain-general assessment of scientific reasoning is hard to achieve but also that the bias due to domain-specific item aspects is not large enough to make group differences disappear. The relation of scientific reasoning with general reasoning needs more attention, since it might be stronger than what was expected so far, considering the results from the bifactor model analysis. The connection with science knowledge might be due to a hidden variable mediating between scientific reasoning and

science knowledge. A weak general scientific reasoning factor seems to exist but a big part of the variance in the CTSR was item-specific and this should always be considered when interpreting scientific reasoning tests. The criterion validity results warrant further exploration of this psychometric property. Furthermore, the observed ceiling effects regarding item difficulty and the low number of items discriminating between higher performing students imply that scientific reasoning tests should be adapted to specific populations regarding their item difficulty. Finally, Study 2 concluded that a focus on the core scientific reasoning skills represented in the F. Fischer et al. (2014) model seems wise, that the deployed methods were useful, and that using these methods or similar ones should become the norm in the evaluation of scientific reasoning tests.

4.2 Theoretical and Methodological Implications

4.2.1 Skills belonging to scientific reasoning

The Introduction to this thesis described how conceptualizations of scientific reasoning differ regarding the number and types of skills they acknowledge. Does this variety of conceptualizations imply that scientific reasoning is "a word with so many meanings that finally it has none" as Spearman (1927, p. 14) once stated about intelligence? According to the review this seems not to be the case, at least not on the basis of the skills included in tests as there are four skills that were tested in more than half of the tests: *hypothesis generation*, *evidence generation*, *evidence evaluation*, and *drawing conclusions*.

Is it reasonable, based on this result, to conclude that scientific reasoning only consists of these four skills and that, similar to the construct of the understanding of the nature of science (NOS), there is a wide agreement about the scientific reasoning construct in its according assessments (Osborne et al., 2003)? Unfortunately, and probably unsurprisingly,

the solution is not that simple. First, it should be noted that even with NOS the case of wide agreement is not that clear (Sinatra & Chinn, 2012). Second, a problem exists in the field of psychometric testing that has long been known, namely "the tendency to measure first that which seemed most feasible" (Herring, 1918, p. 558). It is possible that it is easier to create items for the four commonly tested scientific reasoning skills but that would not necessarily mean that other skills are less important to the construct. When we combine this problem with the results from Study 2 that there is no strong general scientific reasoning factor and that many facets of the construct should be included in the assessment of the whole broad construct, it seems like we should not prematurely conclude that scientific reasoning is just made up of four skills. These four are very central skills of course but, empirically speaking, they should not be the only skills that we consider to be scientific reasoning skills.

Does this mean we are back at the beginning and, just as the construct of intelligence, in danger of a circularity trap, in which scientific reasoning is defined by whatever is measured in a scientific reasoning test (Boring, 1923; Roskam, 1989)? While this is a danger that should not be ignored, the two studies of this thesis exemplify that it is not completely arbitrary which skills can be seen as parts of scientific reasoning. Study 1 showed that quantitative reasoning is part of several scientific reasoning tests but Study 2 showed that quantitative reasoning items are less strongly connected to a general scientific reasoning skill. Quantitative reasoning seems to be a higher level skill just like the model by Shavelson and Huang (2003) suggests. Thus, scientific reasoning conceptualizations without quantitative reasoning, e.g. the conceptualization by F. Fischer et al. (2014), should be preferred.

The studies in this thesis not only allow conclusions about the nature of the scientific reasoning construct itself but of course also about how to assess scientific reasoning. There have been calls in the literature that we need diverse assessments, especially when it comes to assessing competencies in a higher education setting (Shavelson & Huang, 2003; Zlatkin-

Troitschanskaia et al., 2015). These calls for diverse assessments underline the above-mentioned conclusion that scientific reasoning should not only be defined by what is currently measured in the majority of scientific reasoning tests.

In a situation in which we expect a sizeable influence of specific scientific reasoning aspects, the recommendation from Study 1 to compare different scientific reasoning tests becomes especially important. In particular, we need to check more often if tests have a strong connection when they measure the same specific scientific reasoning skill but use different items to do so. If this connection cannot be found it would imply that item-specific elements that are independent from the skill that is measured, e.g. the use of a certain phrase or figure, have a strong influence on the results of scientific reasoning tests. This would limit the generalizations that can be drawn from scientific reasoning tests. It would be unclear if a person with a high score in a test actually mastered the measured scientific reasoning skills in general or if their performance is bound to the specific way the skill was assessed in this test.

Even if we agree that we need to measure all scientific reasoning skills and not only the skills for which items can be easily created this does not answer the question of whether we should try to include all of the skills into a single test or if we should rather construct many different tests that focus each on a single skill. Pellegrino et al. (2014) suggest testing several skills in combination and to even include questions aiming at content knowledge into the same tasks. Of course, this makes testing more complicated, because if you want to include all possible combinations of content knowledge and skills you need (at least) a number of tasks that is the number of content areas by the number of skills. The interpretation of test results also gets more difficult in such a case since it is harder to determine what is actually measured by the test. However, Gustafsson and Åberg-Bengtsson (2010) argue that this additional effort will be worth it. They warn that the results from tests that only measure a single scientific reasoning skill will always be confounded by a general factor that is

unaccounted for. They also argue that bifactor models offer a possibility to adequately describe more complex tests, which can be homogeneous in spite of multiple identifiable influences on the results. The use of bifactor models would be different from the way they were used in Study 2, in which scientific reasoning and three general reasoning scales were treated as subtests of general reasoning. When applied to scientific reasoning in isolation a bifactor model would imply that we should have scientific reasoning tests that measure several skills with several items for each skill. Then, the skills would be treated as subscales of scientific reasoning, which itself would also be represented by a general scientific reasoning factor. This way it is possible to separate the influence of a general scientific reasoning factor from skill-specific factors.

When we conduct an overall evaluation of the skills that are part of scientific reasoning the following picture emerges in this thesis: The Introduction showed that tests that were developed based on Inhelder and Piaget (1958) narrowed down the construct of scientific reasoning in comparison to very early scientific reasoning tests. Then we saw that newer conceptualizations are including more skills again (this is not reflected in the number of skill assessed per test, though). The addition of *formulating questions* to scientific reasoning tests fits the conclusion from Study 2 that all aspects of scientific reasoning should be assessed directly instead of inferring the level of some scientific reasoning skills from a general scientific reasoning factor. Finally, Study 2 showed how a combination of carefully crafted models and evidence from test analysis can help us decide which skills should be part of the conceptualization and the assessment of scientific reasoning.

4.2.2 The domain generality vs. domain specificity of scientific reasoning

While there are domain-general conceptualizations of scientific reasoning, it seems difficult to construct items that adhere to this conceptualization. Study 2 discovered

differences between the epistemologically close domains of physics and biology in a test that approaches scientific reasoning in a rather domain-general way, which probably makes it illusionary to achieve complete domain generality in the assessment of scientific reasoning. Considering this, the trend away from domain-general conceptualizations, which was apparent in the review, is reasonable.

However, a completely domain-specific conceptualization of scientific reasoning would also not be in accordance with the evidence. The physics students did not fail the parts of the CTSR in which the biology students had an advantage. On the contrary, the items were still easier for physics students, so most of them were able to apply their skills to these items.

It becomes clear then that one should be wary if a scientific reasoning test claims to measure scientific reasoning in either a completely domain-general or a completely domain-specific way. It seems that the compromising position in the debate of domain generality vs. domain specificity, which several authors have backed in the past (Erduran, 2007; Karmiloff-Smith, 2012; Klahr & Dunbar, 1988; Niaz, 1995; Zimmerman, 2000), should be preferred instead. In accordance with this position is the idea that there is a theoretical maximum level for the performance of a skill but that this maximum level is not always observable (K. W. Fischer & Farrar, 1987). The skill is not thought of as being tied to a single domain, but there will be performance differences depending on the context the skill is tested in. Thus, we should be careful to make an inference about the degree a student has mastered a skill just from observing one instance of applying the skill. Instead, the performance of the student should be tested in several instances using items with different domain contexts.

The review revealed that implementing the compromising position is not extensively done so far by test authors. It remains an open task to determine the range of the generalization of scientific reasoning skills. There is evidence that the range will be different for different skills. For instance, one study showed that planning investigations was more

generalizable than interpreting results in three tasks in a physics domain (Kok-Auntoh & Woolnough, 1994).

A general conclusion from the studies in this thesis is that it would be helpful for test authors to be more specific about their assumptions about domain generality in the future. Study 2 showed that if we explore new analysis methods there are potential ways to get a grip on this domain generality vs. domain specificity debate from the methodological side. Additionally, Study 2 demonstrated that the methods that were used so far to test the assumptions in this debate at a university level are not sufficient to detect bias.

4.2.3 The factorial structure of scientific reasoning

Regarding the factorial structure of scientific reasoning the results from Study 2 can help explain why Study 1 found scientific reasoning conceptualizations with a varying number of dimensions and that large-scale assessments choose a position that has both unidimensional and multidimensional elements. In Study 2 we saw that the items do have something in common but also that the item-specific parts of the variance were rather high. Thus, Study 2 showed that all conceptualizations were correct in some way: There is a (weak) general scientific reasoning factor but several minor factors are also very important. The strength of the general factor compared to the specific factors needs to be explored in detail but it seems reasonable to suggest at this point that scientific reasoning conceptualizations should consider both aspects.

This complicated structure of scientific reasoning calls for statistical analyses using models that recognize this complexity. Study 1 supports the notion that test authors should conduct these advanced analyses of the factorial structure of their assessments on a routine basis. When doing this, test authors should openly discuss the fit of the model and possible alternative models. They should not insist on a unidimensional structure when there is

evidence against this model as was the case with the CTSR (Lawson, 1978; Pratt & Hacker, 1984). Additionally, when subscales are formed the authors should check if the subscales have a merit beyond a general factor.

4.2.4 Differences to general reasoning and science knowledge

Although Study 2 demonstrated that the connection of scientific reasoning and general reasoning is an issue that should not be ignored, Study 1 unfortunately revealed that this is too often the case. The studies in this thesis strengthen the call to end the disregard for this connection and to include an assessment of general reasoning when evaluating scientific reasoning tests. On the one hand, we can conclude from Study 2 that the view held by Simon (1966) that scientific reasoning has nothing special to offer beyond general reasoning is not completely unreasonable. On the other hand, the regression results show that the two constructs are not identical and the results from the bifactor model analysis indicate that a focus on core scientific reasoning skills might be promising for establishing the difference between the two constructs. A complete independence seems unreasonable, though. The overlap between scientific reasoning and general reasoning makes it hard to embed scientific reasoning in a nomological net (Ziegler & Hagemann, 2015), which should inspire test authors to be as clear as possible when defining the aspects of scientific reasoning they intend to measure.

Of course, future studies might not support the high overlap between general reasoning and scientific reasoning, but even if they do this would not necessarily mean that the assessment of scientific reasoning is obsolete. When researchers manage to make a clear distinction between the two constructs on a theoretical level it is possible to imagine research questions which would require measuring scientific reasoning. For instance, it could be relevant to measure scientific reasoning as an outcome of higher education if we want to

know if students acquire these specific skills. Continuing to measure scientific reasoning would be especially important considering the assumption that it is easier to train scientific reasoning skills than general reasoning. This assumption is backed by evidence showing that scientific reasoning can be trained to a large extent (Engelmann et al., 2016). However, there are also some situations in which we should refrain from measuring scientific reasoning if the high overlap is supported. In studies in which we are mainly interested in predicting a criterion variable, it would be redundant and economically unreasonable to measure both constructs if scientific reasoning does not explain a significant part of the variance beyond general reasoning.

The results from Study 2 about the connection between scientific reasoning and science knowledge support the relevance of domain-specific knowledge. However, it remains unclear from Study 2 which aspects of science knowledge or which intermediary variables are responsible for this connection and if scientific reasoning skills help acquire science knowledge. Thus, in addition to general reasoning, science knowledge tests and possible mediating variables between scientific reasoning and science knowledge should also be included in evaluations of scientific reasoning tests.

4.2.5 Further implications regarding psychometric properties

What else can we learn from the presented studies about the psychometric properties of scientific reasoning tests? The review showed that aspects such as domain generality assumptions, the factorial structure, reliability, divergent validity, convergent validity, and criterion validity are not checked often enough. Study 2 showed that some of these properties, namely reliability, divergent validity, and criterion validity, might not be as high as we would want them to be. Furthermore, the importance of item-specific aspects will complicate comparisons of and generalizations from tests, as was explained in the section about which

skills belong to scientific reasoning. However, Study 2 also demonstrated that a careful interpretation of a scientific reasoning test can lead to interesting and valid findings on the group level, which was one of the conclusions from Study 1. In Study 2 this is true, for instance, regarding the differences between the three majors, which remained significant after controlling for other influences and they would be even bigger if there would be no domain-specific bias.

When we take a look at how the psychometric properties of scientific reasoning tests are checked, the results from both studies imply that test authors should invest more effort and deploy more modern methods. Study 2 emphasized the point by Heene (2007) that using a Rasch model and observing misfit can still lead to interesting insights. Additionally, Study 2 suggested that modern statistical analyses, e.g. DIF analyses and the bifactor model as well as other methods that analyze the validity of interpreting subscale scores, can be valuable tools for the evaluation of scientific reasoning tests.

4.2.6 Test construction

The literature provides various recommendations regarding test construction that seem relevant in the light of the results from this thesis. One important recommendation is to think about how your theoretical attributes and the actual test scores are connected (Borsboom & Mellenbergh, 2007). Thus, test authors should explain how the variance in their latent scientific reasoning construct will lead to variance in the test scores from a theoretical perspective. For instance, an author should be able to state why it makes sense from a theoretical perspective that a person with a high *evidence generation* skill will succeed in a task in which the control-of-variables strategy has to be applied. As was stated in the Discussion of Study 2, authors who try to estimate the difficulty of new scientific reasoning items should consider aspects like the length of answer options, specialist terms, or if test

takers have to work with tabulated data as these are known to have an influence on item difficulty (Stiller et al., 2016). If test authors neglect these aspects they will falsely add the variance in test items that is explained by these aspects towards the variance explained by the construct they intended to measure and thus overestimate the variance explained by the construct. Additionally, they might misjudge the adequacy of the test difficulty for their target population. Furthermore, when one evaluates model fit one should explore occurring misfit, adapt the model accordingly, and try to replicate the structure (Ziegler & Hagemann, 2015). Similarly, Pellegrino et al. (2014) state that the construction process should be iterative, that it needs multiple assessments to get a comprehensive grasp of the assessed area, and that a clearly defined construct has to stand at the beginning of the test development process. They add that so far too much focus has been on surface characteristics such as the test format. The reminder that a test construction process involves multiple steps and replications deserves to be highlighted. When we look at the tests from the review, far too often it seems like tests are constructed rather fast and the first analysis has to be instantly perfect. However, this is an unrealistic expectation that researchers should abandon in the future.

These recommendations from the literature are complemented by results from the thesis: It is important to remember the conclusion from the review that we should first try to learn from analyses of current tests. Then, when a new test is constructed, it should avoid the mistakes of the past. The review is also a good guide to areas of scientific reasoning assessment that are not covered enough so far. Additionally, Study 2 names aspects that are crucial in the test construction and evaluation process. For instance, assumptions along with ways to test them should be stated clearly at the beginning. Additionally, a test should always include several items for each scientific reasoning skill that is measured in order to check if the items measure the intended skill or mainly contain item-specific variance.

4.3 Practical Implications

4.3.1 Test selection

This section gives an overview of aspects of the presented studies that are useful for people who are looking for a scientific reasoning test. It is clear from the review that tests with various test formats and for different target groups exist. However, populations outside of educational institutions are underserved so far. It might be hard to find a fitting test for these populations, especially since Study 2 warned us not to use a test within a population that it was not intended for. It is possible, for instance, that the item difficulties will not be in the optimal range to differentiate between test takers on higher and lower latent ability levels. Thus, the fit of the test to a specific population should definitely be checked before applying the test to a large sample. If these checks indicate a misfit it might be necessary to develop a test that is adapted to the target population.

Once a person looking for a test got an overview of potential tests that could be used, how should they go about selecting one or several tests? The review contains a useful guide for how to proceed in such a situation and Study 2 emphasizes two of the central points of these recommendations. First, it is very important that the selected test really matches the scientific reasoning construct or the subpart of scientific reasoning that one intends to measure. Study 2 demonstrated that minor item differences can matter. Therefore, not any scientific reasoning test will do when you want to measure a specific aspect of scientific reasoning and if you want to measure scientific reasoning as a broad construct the test should not only focus on one scientific reasoning skill, respectively. Similarly, it should be checked if the test measures aspects that are not part of your scientific reasoning conceptualization, e.g. quantitative reasoning. Tests or subscales of tests that do that should be avoided.

Second, it is necessary to find out which psychometric properties of the test were evaluated by the creators of the test as well as by other authors who used the test and what the results were. Study 2 hints at some of the most pressing questions that should be answered by the literature about the test: Can you interpret subscale scores you are interested in or does only the total score carry valid information? Are comparisons between groups that you want to look at in your study free from bias, e.g. the comparison of students from different school types? Does the test tell you anything about aspects of criterion validity that are relevant to you, e.g. predicting science grades? If the test authors have not made these checks yet, try to conduct them before using the test in the final study. Furthermore, it should not be forgotten to have a critical look at the test construction process in addition to the evaluation process. For instance, it is crucial to find out which assumptions about scientific reasoning were made during test construction.

4.3.2 Basing decisions on a scientific reasoning test

Next to matters of test selection and administration another relevant aspect to think about are the decisions that can be based on the test results. In the field of secondary education it was possible to observe the influence of the outcomes of the PISA assessment (OECD, 2007) on educational decision making in recent years. Similar assessments for higher education are in demand and in development at the moment (Shavelson & Huang, 2003; Zlatkin-Troitschanskaia et al., 2015). There are states in the USA where it is required that scientific reasoning is tested in higher education institutions and these states base academic program decisions on the outcomes (State Council of Higher Education for Virginia, 2007).

However, the presented studies remind us of the limitations of the current state of scientific reasoning assessment and thus test results should be interpreted very carefully.

Using another parallel to intelligence testing, we have to avoid a situation where "consequent action affecting the welfare of thousands of persons is proposed, and even taken, on the grounds of – nobody knows what" (Spearman, 1927, p. 15). Despite the variety of tests that exist, and even when a test is selected carefully, it still seems wise to not base high-stake decisions on scientific reasoning tests at the moment. When it comes to higher education, Study 2 shows that it might be unfair to compare universities that vary in something as seemingly minor as the number of biology students. While the bias was not big enough to distort the rank order of group means in the comparison of biology and physics students, this might not be the case when biology is compared with other majors.

Selecting people for graduate programs based on a scientific reasoning test would be especially problematic. Even if measurement invariance holds, it is still possible that there will be a violation of selection invariance among groups with a different latent ability (Borsboom et al., 2008). Selection invariance is important in a situation in which a test is used for selecting qualified people. It means that there is no significant difference for any subgroup regarding the percentage of qualified people who are correctly identified as qualified (sensitivity) and the percentage of unqualified people who are correctly identified as such (specificity). If selection invariance is violated, a test will have a different sensitivity and specificity for different groups.

Many scientific reasoning tests are not based on a model from item response theory (IRT), which should be required if high-stake decisions are made (Heene, 2007). Additionally, when test results start to have severe consequences, the well-known problem arises that too much focus within the education system will be placed on improving test scores as opposed to improving instruction (Wiliam, 2001). In a situation in which the quality of the tests is not fully explored, which is the case for the assessment of scientific reasoning, the damages resulting from this problem would probably be even worse. Thus, it would be

good to listen to the people who remind us that it will take a long time before educational standards can lead to high-stake assessments (Pellegrino, 2013). In the meantime, we should not forget that on top of summative assessment it is also worthwhile to invest in formative assessment (Wiliam, 2011).

4.4 Limitations

The conclusions that one can draw from the results of the presented studies are limited by several factors: Both studies focused exclusively on scientific reasoning and excluded NOS and scientific argumentation. While it is a good idea to not confound these different constructs, the focus on scientific reasoning naturally limits the scope of this thesis. Furthermore, the review did not evaluate the results of psychometric property checks in depth and instead just reported how psychometric properties were checked because the results were too different in nature to make a valid comparison. Besides, there might be tests that target scientific reasoning skills but use none of the terms that were used in the literature search when they describe what they are measuring. These tests would not have had a chance to be included in the review. However, our selection of tests should reflect the set of tests that a person who is looking for a scientific reasoning test would come across as long as they use the common search terms for scientific reasoning skills. The selection of tests for the review was further narrowed by the requirement that the test has to provide at least some form of validity check. While it might be insightful to look at tests without validity checks, this criterion was necessary to exclude any ad-hoc measurements, which were not meant for repeated use. Finally, considering that the second wave of test development probably has not reached its end, it seems advisable to repeat or update the review of scientific reasoning tests within a few years.

Study 2 only looked at university students and only used a single scientific reasoning test, albeit a very commonly used one. This reduces the width of generalizations that can be drawn from the study. The observation that item-specific aspects play such a big role further diminishes the generalizability of the results. It makes it more likely that different results will be obtained with different items. However, the importance of minor factors is also an interesting result in and of itself. Additionally, the conclusions that can be drawn from Study 2 are somewhat limited by ceiling effects in the scales. The ceiling effects might have caused a reduction of variance which would make it more difficult to detect the accurate size of the relations between variables.

4.5 Suggestions for Future Research

Study 2 built on the suggestions of Study 1 to evaluate current scientific reasoning tests in detail. While the presented studies thus contributed to the discussion surrounding scientific reasoning and its assessment, there are of course many open questions and problems left. This section will explore various avenues for future research.

4.5.1 Improving the conceptualization of scientific reasoning

On a conceptual level we not only have to think about which skills are part of scientific reasoning but also how a correct and realistic application of a skill would look like in order to construct adequate test items. For instance, correct answers in scientific reasoning tests often require that a hypothesis is abandoned based on a single piece of evidence. It might be more accurate, though, to assume a Bayesian approach in which multiple pieces of evidence have to accumulate over time before a hypothesis is changed (Gopnik, 2012; Szu & Osborne, 2011). If more research shows that people act in this way it should be reflected in test items

in the future. In a similar fashion, current tests were criticized because common items about the control of variables strategy only focus on the effects of a single independent variable, while in real world scenarios multiple influences have to be controlled (Kuhn, Iordanou, Pease, & Wirkala, 2008). So far, research on how to assess these more complex scientific reasoning situations is lacking.

4.5.2 Evaluating criterion validity

In the Introduction it was mentioned that some see mastering scientific reasoning as a merit of its own (Harlen, 1999). For everyone else, the criterion validity of scientific reasoning tests is a crucial issue. Before we can gather evidence about the criterion validity of a test, we have to agree on the criterion that should be used. Research about argumentation skills showed that these skills help acquire content knowledge through elaborating one's own explanation, learning from others, and giving you a stronger reason to believe a claim (Chinn & Clark, 2013). Support for learning content knowledge would also be an interesting criterion for scientific reasoning tests. There is research showing that it is indeed a promising approach to observe how scientific reasoning skills support the acquisition of new knowledge (Chen & Klahr, 1999; Edelsbrunner, Schalk, Schumacher, & Stern, 2015). Thus, it seems wise to adopt the advice by Wecker, Hetmanek, and Fischer (2016) to establish the simultaneous training of domain-general competencies and domain-specific knowledge as its own research topic. A recent meta-analysis showed that scientific reasoning skills can be trained (Engelmann et al., 2016) but the seemingly high overlap with general reasoning might indicate that there is a limit to the effects of such trainings. Another aspect of criterion validity that needs further research is the claim that skills such as *formulating questions* or *evidence evaluation* are important for civic participation (Rudolph & Horibe, 2016). Claims like this are the foundation on which the importance of the assessment of scientific reasoning

is grounded, so it is necessary to find evidence for a connection between scientific reasoning and measures of civic participation in the future.

4.5.3 Exploring new test formats

Study 1 revealed that the variety of test formats is growing in newer tests. Two of the tests used automated analyses of simulated experiments that were conducted within a software program (Gobert et al., 2013; Su et al., 2011). A next step could be to integrate the assessment of scientific reasoning into virtual 3D environments, an approach that seems promising (Clarke-Midura, Code, Zap, & Dede, 2012; Hickey, Ingram-Goble, & Jameson, 2009). Such tests could satisfy the calls for performance assessments (Shavelson, Baxter, & Pine, 1991), embedding assessments into games (Shute & Kim, 2013), and the integration of new technologies into assessments (Tucker, 2009) all at once. A possible advantage could be that these tests generate more data about why students fail in scientific reasoning tasks. However, a more important task than changing these surface characteristics might be to make scientific reasoning assessments more construct-driven (Kind, 2013). Of course, these potentially new tests also need to meet psychometric standards. The use of modern technology will not automatically lead to better tests. Apart from new test formats, it is necessary to develop tests that cover scientific reasoning skills (e.g. *problem identification* or *questioning*) and target groups (e.g. adults who finished their formal education) that were neglected so far.

4.5.4 Inspirations by other areas of assessment

It is true that the assessment of scientific reasoning faces many challenges. However, scientific reasoning is not the only research field that faces these challenges. They are shared by other psychological constructs. Thus, in the following paragraphs we will look at several

of these constructs, the challenges the according assessments face, and the lessons we can learn from these research fields for the assessment of scientific reasoning.

First up is the measurement of basic Newtonian concepts, a central content knowledge area from physics. The most common test in this field is the Force Concept Inventory (FCI, Hestenes et al., 1992). As is the case with many scientific reasoning tests, its factorial structure is not as simple as the original test authors thought (Huffman & Heller, 1995; Scott, Schumayer, & Gray, 2012). Additionally, the test was checked for its connection with scientific reasoning, using the Lawson test (Coletta & Phillips, 2005). Authors also conducted studies to check its reliability (Lasry, Rosenfield, Dedic, Dahan, & Reshef, 2011) and its connection with SAT scores and knowledge of other content areas (Diff & Tache, 2007). Furthermore, several authors used various IRT methods to test the attractiveness of item distractors, the dimensionality of the test, and its measurement invariance (Morris et al., 2012; Planinic, Ivanjek, & Susac, 2010; Scott & Schumayer, 2015).

It is noteworthy how thoroughly the test was checked by a lot of different authors over an extended period of time. The assessment of scientific reasoning could benefit from adopting a similar approach to its most well-known tests. Unfortunately, it is common that only the original test authors try to validate a test with a very limited amount of studies.

The second construct of interest is problem-solving. There are many parallels between scientific reasoning and problem-solving: Problem-solving is also considered a 21st century skill and important for the workforce, many conceptualizations with a different set of subskills exist for problem-solving, despite being a complex construct the according assessments often have a rather simple structure, assessments are often created ad-hoc and not grounded in a theory, the validity of problem-solving assessments is not checked often enough, e.g. comparisons with other problem-solving tests are rarely made and the difference to general reasoning is not established yet, and there is a debate whether problem-solving is

domain-general or domain-specific (Adams & Wieman, 2015; Greiff, 2012). There are two interesting approaches of how to handle these challenges that will be pointed out here.

Adams and Wieman (2015) analyzed a task involving a complete problem-solving process in great detail via think aloud interviews. All together, they discovered 44 subskills that were necessary to solve the problem and they included knowledge aspects in this list. The authors argue that problem-solving is hard to separate from content knowledge, assessments should focus on evaluating subskills, and instruction should also happen on the subskill level. Since Study 2 showed that many item-specific aspects were important in solving scientific reasoning tasks, it might be helpful to conduct similarly detailed analyses with scientific reasoning tests. One study, which did use think aloud procedures with a scientific reasoning test, discovered that prior beliefs influenced some answers and that correct answers could sometimes be found via construct-irrelevant strategies (Thelk & Hoole, 2006).

The second approach suggests to separate problem-solving phases artificially in the assessment process and to exclude knowledge from items as much as possible (Greiff, 2012). Although this General Discussion presented arguments why this might not be a good idea (Gustafsson & Åberg-Bengtsson, 2010; Pellegrino et al., 2014), the evidence is far from being conclusive at this point. If it turns out that the approach by Greiff (2012) advances the assessment of problem-solving, we can transfer the learned lessons to the assessment of scientific reasoning. The call from Greiff (2012) to better connect theories and assessments is definitely a recommendation to which the field of scientific reasoning assessment should listen.

The third and last field we will look at is the assessment of depression. This might seem strange at first but there are interesting parallels between the assessment of scientific

reasoning and the assessment of depression. In the latter field some recent developments happened that deserve the attention of researchers from the former field.

In the last century almost 300 scales aiming to measure depression were developed (Fried, 2016). Most of what is known about depression is based on tests that are at least four decades old (Fried, Epskamp, Nesse, Tuerlinckx, & Borsboom, 2016), which is not completely unlike the situation of scientific reasoning and its assessment given e.g. that the CTSR is still in use (Bao et al., 2009). It is a common practice in the assessment of depression to use a sum score to represent one underlying construct. This is done despite the fact that measurement invariance does not hold over time and a unidimensional model is not an accurate description of the data because apart from a strong first factor several other minor factors are needed to explain the covariance of items (Fried, van Borkulo, et al., 2016). These aspects are also all true for scientific reasoning. An especially interesting aspect of this last study, which was conducted with 3509 depressed patients, is that the dimensionality of assessments was higher before a treatment compared to the end of the treatment period. This means the number of factors that were found in a factor analysis decreased from the pretest to the posttest. It would be interesting to observe if this is also true for measurements used in the context of scientific reasoning trainings. It would implicate that the tests are measuring something else before and after an intervention. Two suggestions that the authors offer in this study are to either construct assessments that focus on single symptoms or to use formative models. In a formative model, indicators of the latent construct are no longer exchangeable and the sum score does not represent a latent variable any longer but becomes a mere index instead, which stands for whatever is measured in the specific test (Edwards & Bagozzi, 2000). Both suggestions are not without problems. The problem with constructing assessments that focus on a single aspect – the results will always be confounded by the influence of a general factor and thus do not only represent this single aspect – was

mentioned before in this General Discussion (Gustafsson & Åberg-Bengtsson, 2010). Formative models also have some fundamental disadvantages, as Edwards and Bagozzi (2000) explain: It is not possible to make causal statements with formative models and the indicators of the latent construct are assumed to be free of measurement error, which is not realistic. Furthermore, it is often neglected that not all the variance of the construct can be explained by the indicators. Thus, we should be careful to adopt such an approach for scientific reasoning even if it might seem worthwhile to consider this option when more evidence is generated that scientific reasoning items are not interchangeable indicators of a latent construct.

Since both prior suggestions are problematic, a third one might be more interesting. This suggestion can be found in a study by Fried, Epskamp, et al. (2016), in which they used a network approach to model depression. Using this approach, depression is conceptualized as a web of causally connected symptoms instead of one underlying latent variable. For instance, insomnia could cause fatigue which in turn could cause concentration and psychomotor problems. The model allows for feedback loops, consists of symptoms from different depression scales, and also includes symptoms from neighboring constructs like anxiety. This model was then applied to the data from 3463 patients. The authors were interested in the centrality of depression symptoms, i.e. how connected they are with other symptoms, since not all symptoms are equally good indicators of depression. Central symptoms activate more other symptoms when they are activated. The authors reason that it would be good to pay attention to central symptoms for diagnosing depression while also including more symptoms in general because the symptoms are not interchangeable indicators of depression. The results revealed that some central symptoms were expected to be central but there were also other central symptoms that are commonly associated with anxiety. Depression and anxiety apparently do not form distinct symptom clusters and thus

should be assessed together. Additionally, the network model is potentially useful for planning interventions. Symptoms with high centrality could be targeted with specific interventions to stop the spread of the activation.

What are the lessons that the assessment of scientific reasoning can learn from this approach? First, it might be worthwhile to conceptualize scientific reasoning skills in a network model instead of a latent variable model. It is possible to imagine, for instance, that *hypothesis generation* influences *evidence generation*, which in turn influences *evidence evaluation* and *drawing conclusions*, and finally a feedback process might occur in which a conclusion leads to *hypothesis generation* again. Such a conceptualization seems not too unreasonable since it would be similar to the already used problem-solving approaches of scientific reasoning. Second, the idea to include a high number of scientific reasoning skills from different assessments into the network and to give special attention to central skills instead of simply trusting a sum score would fit the results about item-specific factors from Study 2. Third, there are constructs that are conceptually close to scientific reasoning, like NOS, and it would be possible to include items from these constructs into the network to see how distinct the constructs really are. Fourth, a network model, if it can be established, might be useful for the training of scientific reasoning. It would provide recommendations about which central skill should be trained first, since they would activate many other skills and thus increase the training efficiency. The last noteworthy aspect of network models is that they have proven to be useful in more than one field already: Researchers also already used similar approaches to model intelligence (van der Maas et al., 2006).

One final parallel between scientific reasoning and depression deserves to be mentioned. Fried (2016) describes how the responsiveness of a depression scale is considered to be an important psychometric property. A high responsiveness means that a scale is sensitive to improvements in patients who received a therapy. The idea of responsiveness is

based on the assumption that the most responsive scales provide the most valid measures of the improvement. However, Fried (2016) explains that three more basic assumptions have to be met for this to be true: unidimensionality, temporal invariance (which is measurement invariance over time), and content validity. It was already mentioned that depression scales are neither unidimensional, nor possess measurement invariance over time. Despite this, a valid interpretation of responsiveness as an indicator of superior assessments might still be possible in a formative model. However, the depression scales are so different in content that it would be hard to compare them at all in such a formative model. It is especially problematic to evaluate short scales based on their responsiveness. Short scales do not represent the complete depression construct but instead might have been tailored to detect certain aspects of alleviating a depression that the specific intervention is particularly good at. Thus, a scale measuring a broader range of symptoms but with a lower responsiveness might give us a more accurate picture of the effects of an intervention on the complete depression construct.

A problem that is similar to the issue of responsiveness to therapies can be observed with the assessment of scientific reasoning. Several authors tried to establish the validity of their test by showing the responsiveness of the test to a scientific reasoning training (Klos et al., 2008; Molitor & George, 1976; Ross & Maynes, 1983; Shaw, 1983). However, considering the results of Study 2 it is not unlikely that the scales are not fulfilling the necessary requirements of unidimensionality and measurement invariance to make this conclusion. Not all scientific reasoning tests measure the same scientific reasoning skills so it is possible that a responsive scientific reasoning test simply measures the skills that are conceptually closest to the contents of the training without representing scientific reasoning in general. Therefore a test validation strategy based on responsiveness should be avoided. Of

course, once the validity of a test has been established it can be used to evaluate interventions that it was not tailored for.

To summarize, several other assessment fields face challenges that are similar to the challenges of scientific reasoning assessment. Researchers interested in the assessment of scientific reasoning should feel encouraged to try out the solutions that were suggested in these other research fields. Once we have learned more about scientific reasoning from current tests, we can move on to create better tests.

4.6 In Closing

About 26 years ago, R. E. Mayer and Duemler (1990) emphasized three aspects of scientific reasoning that researchers should focus on: the clarity of theories, the definition of constructs, and the validity of research methods. These three aspects remain important today and when we only consider the last few years it might seem like no progress was made. However, if the whole 100 years of its development are considered, it becomes clear that the assessment of scientific reasoning has advanced. Answer options are less subjective, tests rely less on verbal reasoning and in the last decades the field started to move away from unrealistic assumptions about complete domain generality and a singular scientific reasoning dimension. The presented studies aim to support this positive development. They help clarify the concept of scientific reasoning, point out problems, which makes it easier to tackle them, and present ways to reduce these problems. Thus, the studies contribute to improving the assessment of scientific reasoning, which is the task that Osborne (2013) called the 21st century challenge for science education.

References

Abd-El-Khalick, F. (2012). Examining the sources for our understandings about science: Enduring conflations and critical issues in research on nature of science in science education. *International Journal of Science Education*, *34*(3), 353–374. https://doi.org/10.1080/09500693.2011.629013

Abd-El-Khalick, F., BouJaoude, S., Duschl, R., Lederman, N. G., Mamlok-Naaman, R., Hofstein, A., … Tuan, H. (2004). Inquiry in science education: International perspectives. *Science Education*, *88*(3), 397–419. https://doi.org/10.1002/sce.10118

Adams, W. K., & Wieman, C. E. (2015). Analyzing the many skills involved in solving complex physics problems. *American Journal of Physics*, *83*(5), 459–467. https://doi.org/10.1119/1.4913923

Alexander, P. A., Jetton, T. L., & Kulikowich, J. M. (1995). Interrelationship of knowledge, interest, and recall: Assessing a model of domain learning. *Journal of Educational Psychology*, *87*(4), 559–575.

Amthauer, R., Brocke, B., Liepmann, D., & Beauducel, A. (2001). *Intelligenz-Struktur-Test 2000 R*. Göttingen: Hogrefe.

Anderson, J. R. (1982). Acquisition of cognitive skill. *Psychological Review*, *89*(4), 369–403.

Anderson, J. R. (1993). *Rules of the mind*. Hillsdale, NJ: Lawrence Erlbaum Associates.

Anderson, J. R. (1996). ACT: A simple theory of complex cognition. *American Psychologist*, *51*(4), 355–365.

Anderson, J. R., & Schunn, C. D. (2000). Implications of the ACT-R learning theory: No magic bullets. In R. Glaser (Ed.), *Advances in instructional psychology: Educational design and cognitive science, vol. 5* (pp. 1–34). Mahwah, NJ: Lawrence Erlbaum Associates.

Asendorpf, J. B. (2011). *Persönlichkeitspsychologie - für Bachelor* (2., überarbeitete und aktualisierte Auflage). Heidelberg: Springer.

Azarpira, N., Amini, M., Kojuri, J., Pasalar, P., Soleimani, M., Khani, S. H., … Lankarani, K. B. (2012). Assessment of scientific thinking in basic science in the Iranian second national olympiad. *BMC Research Notes*, *5*(61), 1–7.

Baird, W. E. (1989). Correlates of student performance in the science Olympiad: the Test of Integrated Process Skills and other variables. Paper presented at the Annual Meeting of the National Association for Research in Science Teaching, San Francisco, CA.

Bao, L., Cai, T., Koenig, K., Fang, K., Han, J., Wang, J., … Luo, Y. (2009). Learning and scientific reasoning. *Science*, *323*(5914), 586–587.

Bauer, D. J. (2016). A more general model for testing measurement invariance and differential item functioning. *Psychological Methods*. https://doi.org/10.1037/met0000077

Bauer, H. H. (1994). *Scientific literacy and the myth of the scientific method*. Champaign, IL: University of Illinois Press.

Bayazit, A., & Aşkar, P. (2012). Performance and duration differences between online and paper–pencil tests. *Asia Pacific Education Review*, *13*(2), 219–226. https://doi.org/10.1007/s12564-011-9190-9

Berger, M. (2016). DIFtree: Item focused trees for the identification of items in differential item functioning (Version 2.0.4) [R package]. Retrieved from https://CRAN.R-project.org/package=DIFtree

Berland, L. K., & McNeill, K. L. (2010). A learning progression for scientific argumentation: Understanding student work and designing supportive instructional contexts. *Science Education*, *94*(5), 765–793. https://doi.org/10.1002/sce.20402

Blair, G. M. (1940). The validity of the Noll test of scientific thinking. *Journal of Educational Psychology*, *31*(1), 53–59.

Blömeke, S., Gustafsson, J.-E., & Shavelson, R. J. (2015). Beyond dichotomies: Competence viewed as a continuum. *Zeitschrift Für Psychologie*, *223*(1), 3–13. https://doi.org/10.1027/2151-2604/a000194

Boone, W. J., Staver, J. R., & Yale, M. S. (2014). *Rasch analysis in the human sciences*. Dordrecht: Springer.

Boring, E. G. (1923). Intelligence as the tests test it. *The New Republic*, *10*(23), 35–37.

Borsboom, D. (2006). When does measurement invariance matter? *Medical Care*, *44*(11), 176–181.

Borsboom, D., & Mellenbergh, G. J. (2007). Test validity in cognitive assessment. In J. P. Leighton & M. J. Gierl (Eds.), *Cognitive diagnostic assessment for education: Theory and applications* (pp. 85–115). New York, NY: Cambridge University Press.

Borsboom, D., Mellenbergh, G. J., & Van Heerden, J. (2002). Different kinds of DIF: A distinction between absolute and relative forms of measurement invariance and bias. *Applied Psychological Measurement*, *26*(4), 433–450. https://doi.org/10.1177/014662102237798

Borsboom, D., Mellenbergh, G. J., & van Heerden, J. (2004). The concept of validity. *Psychological Review*, *111*(4), 1061–1071. https://doi.org/10.1037/0033-295X.111.4.1061

Borsboom, D., Romeijn, J.-W., & Wicherts, J. M. (2008). Measurement invariance versus selection invariance: Is fair selection possible? *Psychological Methods*, *13*(2), 75–98. https://doi.org/10.1037/1082-989X.13.2.75

Bridewell, W., Sánchez, J. N., Langley, P., & Billman, D. (2006). An interactive environment for the modeling and discovery of scientific knowledge. *International Journal of Human-Computer Studies*, *64*(11), 1099–1114. https://doi.org/10.1016/j.ijhcs.2006.06.006

Brigandt, I. (2010). Scientific reasoning is material inference: Combining confirmation, discovery, and explanation. *International Studies in the Philosophy of Science*, *24*(1), 31–43. https://doi.org/10.1080/02698590903467101

Brown, N. J. S., Nagashima, S. O., Fu, A., Timms, M., & Wilson, M. (2010). A framework for analyzing scientific reasoning in assessments. *Educational Assessment*, *15*(3–4), 142–174. https://doi.org/10.1080/10627197.2010.530562

Bühner, M. (2011). *Einführung in die Test- und Fragebogenkonstruktion.* München: Pearson Studium.

Bühner, M., & Ziegler, M. (2009). *Statistik für Psychologen und Sozialwissenschaftler*. München: Pearson.

Burns, J. C., Okey, J. R., & Wise, K. C. (1985). Development of an integrated process skill test: TIPS II. *Journal of Research in Science Teaching*, *22*(2), 169–177.

Bybee, R., McCrae, B., & Laurie, R. (2009). PISA 2006: An assessment of scientific literacy. *Journal of Research in Science Teaching*, *46*(8), 865–883. https://doi.org/10.1002/tea.20333

Cattell, R. B. (1971). *Abilities: Their structure, growth, and action.* Oxford: Houghton Mifflin.

Cavagnetto, A. R. (2010). Argument to foster scientific literacy: A review of argument interventions in K–12 science contexts. *Review of Educational Research*, *80*(3), 336–371.

Chang, C.-Y., Yeh, T., & Barufaldi, J. P. (2010). The positive and negative effects of science concept tests on student conceptual understanding. *International Journal of Science Education*, *32*(2), 265–282. https://doi.org/10.1080/09500690802650055

Chang, H.-P., Chen, C.-C., Guo, G.-J., Cheng, Y.-J., Lin, C.-Y., & Jen, T.-H. (2011). The development of a competence scale for learning science: Inquiry and communication. *International Journal of Science and Mathematics Education*, *9*(5), 1213–1233.

Chen, Z., & Klahr, D. (1999). All other things being equal: Acquisition and transfer of the control of variables strategy. *Child Development*, *70*(5), 1098–1120.

Chinn, C. A. (2006). Learning to argue. In A. M. O'Donnell, C. E. Hmelo-Silver, & G. Erkens (Eds.), *Collaborative learning, reasoning, and technology* (pp. 355–383). Mahwah, NJ: Lawrence Erlbaum Associates Publishers.

Chinn, C. A., & Clark, D. B. (2013). Learning through collaborative argumentation. In C. E. Hmelo-Silver, C. A. Chinn, C. K. K. Chan, & A. M. O'Donnell (Eds.), *The international handbook of collaborative learning* (pp. 314–332). New York, NY: Routledge/Taylor & Francis Group.

Clarke-Midura, J., Code, J., Zap, N., & Dede, C. (2012). Assessing science inquiry: A case study of the Virtual Performance Assessment project. In L. Lennex & K. F. Nettleton (Eds.), *Cases on inquiry through instructional technology in math and science: Systemic approaches* (pp. 138–164). New York, NY: IGI Publishing.

Cloonan, C. A., & Hutchinson, J. S. (2011). A chemistry concept reasoning test. *Chemistry Education Research and Practice*, *12*(2), 205–209. https://doi.org/10.1039/c1rp90025k

Coletta, V. P., & Phillips, J. A. (2005). Interpreting FCI scores: Normalized gain, preinstruction scores, and scientific reasoning ability. *American Journal of Physics*, *73*(12), 1172–1182. https://doi.org/10.1119/1.2117109

Day, H., & Matthews, D. (2008). Do large-scale exams adequately assess inquiry? An evaluation of the alignment of the inquiry behaviors in New York State's "Living environment regents examination" to the NYS inquiry standard. *American Biology Teacher*, *70*(6), 336–341.

Dejonckheere, P. J. N., Van de Keere, K., Tallir, I., & Vervaet, S. (2013). Primary school science: Implementation of domain-general strategies into teaching didactics. *The Australian Educational Researcher*, *40*(5), 583–614. https://doi.org/10.1007/s13384-013-0119-7

Diff, K., & Tache, N. (2007). From FCI to CSEM to Lawson Test: A report on data collected at a community college. In L. Hsu, C. Henderson, & L. McCullough (Eds.), *AIP Conference Proceedings* (Vol. 951, pp. 85–87). Retrieved from http://link.aip.org/link/?APCPCS/951/85/1

Dillashaw, F. G., & Okey, J. R. (1980). Test of the integrated science process skills for secondary science students. *Science Education*, *64*(5), 601–608.

diSessa, A. A. (2008). A "theory bite" on the meaning of scientific inquiry: A companion to Kuhn and Pease. *Cognition and Instruction*, *26*(4), 560–566. https://doi.org/10.1080/07370000802391760

Donovan, J., Hutton, P., Lennon, M., O'Connor, G., & Morrissey, N. (2008). *National Assessment Program--science literacy year 6 school release materials, 2006* (p. 136). Carlton, South Victoria: Ministerial Council on Education, Employment, Training and Youth Affairs.

Donovan, J., Lennon, M., O'Connor, G., & Morrissey, N. (2008). *National Assessment Program--science literacy year 6 report, 2006* (p. 112). Carlton, South Victoria: Ministerial Council on Education, Employment, Training and Youth Affairs.

Edelsbrunner, P. A., Schalk, L., Schumacher, R., & Stern, E. (2015). Pathways of conceptual change: Investigating the influence of experimentation skills on conceptual knowledge development in early science education. In D. C. Noelle, R. Dale, A. S. Warlaumont, J. Yoshimi, T. Matlock, C. D. Jennings, & Maglio (Eds.), *Proceedings of the 37th annual meeting of the Cognitive Science Society* (pp. 620–625).

Edwards, J. R., & Bagozzi, R. P. (2000). On the nature and direction of relationships between constructs and measures. *Psychological Methods*, *5*(2), 155–174.

Engelhart, M. D., & Lewis, H. B. (1941). An attempt to measure scientific thinking. *Educational and Psychological Measurement*, *1*(1), 289–293. https://doi.org/10.1177/001316444100100123

Engelmann, K., Neuhaus, B. J., & Fischer, F. (2016). Fostering scientific reasoning in education – Meta-analytic evidence from intervention studies. *Educational Research and Evaluation*, 1–17. https://doi.org/10.1080/13803611.2016.1240089

Eraña, Á., & Martínez, S. F. (2004). The heuristic structure of scientific knowledge. *Journal of Cognition and Culture*, *4*(3), 701–729.

Erduran, S. (2007). Breaking the law: Promoting domain-specificity in chemical education in the context of arguing about the periodic law. *Foundations of Chemistry*, *9*(3), 247–263. https://doi.org/10.1007/s10698-007-9036-z

Feyzioglu, B., Demirdag, B., Akyildiz, M., & Altun, E. (2012). Developing a science process skills test for secondary students: Validity and reliability study. *Educational Sciences: Theory and Practice*, *12*(3), 1899–1906.

Fischer, F., Kollar, I., Ufer, S., Sodian, B., Hussmann, H., Pekrun, R., … Eberle, J. (2014). Scientific reasoning and argumentation: Advancing an interdisciplinary research agenda in education. *Frontline Learning Research*, *2*(3), 28–45.

Fischer, K. W. (1980). A theory of cognitive development: The control and construction of hierarchies of skills. *Psychological Review*, *87*(6), 477–531.

Fischer, K. W., & Farrar, M. J. (1987). Generalizations about generalization: How a theory of skill development explains both generality and specificity. *International Journal of Psychology*, *22*(5–6), 643–677.

Fisher, W. P., Jr. (2004). Meaning and method in the social sciences. *Human Studies*, *27*(4), 429–454.

Fitts, P. M., & Posner, M. I. (1967). *Human performance.* Belmont, CA: Wadsworth.

Fraser, B. J. (1979). *Test Of Enquiry Skills [and] handbook*. Hawthorn, Victoria: Australian Council for Educational Research.

Fraser, B. J. (1980). Development and validation of a test of enquiry skills. *Journal of Research in Science Teaching*, *17*(1), 7–16.

Frederiksen, N., & Ward, W. C. (1978). Measures for the study of creativity in scientific problem-solving. *Applied Psychological Measurement*, *2*(1), 1–24.

Frey, A., & Seitz, N.-N. (2009). Multidimensional adaptive testing in educational and psychological measurement: Current state and future challenges. *Studies in Educational Evaluation*, *35*(2–3), 89–94. https://doi.org/10.1016/j.stueduc.2009.10.007

Fried, E. I. (2016). Are more responsive depression scales really superior depression scales? *Journal of Clinical Epidemiology*. https://doi.org/10.1016/j.jclinepi.2016.05.004

Fried, E. I., Epskamp, S., Nesse, R. M., Tuerlinckx, F., & Borsboom, D. (2016). What are "good" depression symptoms? Comparing the centrality of DSM and non-DSM

symptoms of depression in a network analysis. *Journal of Affective Disorders*, *189*, 314–320. https://doi.org/10.1016/j.jad.2015.09.005

Fried, E. I., van Borkulo, C. D., Epskamp, S., Schoevers, R. A., Tuerlinckx, F., & Borsboom, D. (2016). Measuring depression over time . . . or not? Lack of unidimensionality and longitudinal measurement invariance in four common rating scales of depression. *Psychological Assessment*, *28*(11), 1354–1367. https://doi.org/10.1037/pas0000275

Gelman, R., & Brenneman, K. (2004). Science learning pathways for young children. *Early Childhood Research Quarterly*, *19*(1), 150–158. https://doi.org/10.1016/j.ecresq.2004.01.009

Germann, P. J. (1989). The Processes of Biological Investigations Test. *Journal of Research in Science Teaching*, *26*(7), 609–625.

Glaser, R., Schauble, L., Raghavan, K., & Zeitz, C. (1992). Scientific reasoning across different domains. In E. de Corte, M. Linn, H. Mandl, & L. Verschaffel (Eds.), *Computer-based learning environments and problem solving* (pp. 345–373). Berlin: Springer.

Gobert, J. D., Sao Pedro, M., Raziuddin, J., & Baker, R. S. (2013). From log files to assessment metrics: Measuring students' science inquiry skills using educational data mining. *Journal of the Learning Sciences*, *22*(4), 521–563. https://doi.org/10.1080/10508406.2013.837391

Gopnik, A. (2012). Scientific thinking in young children: Theoretical advances, empirical research, and policy implications. *Science*, *337*(6102), 1623–1627. https://doi.org/10.1126/science.1223416

Gormally, C., Brickman, P., & Lutz, M. (2012). Developing a test of scientific literacy skills (TOSLS): Measuring undergraduates' evaluation of scientific information and

arguments. *Cell Biology Education*, *11*(4), 364–377. https://doi.org/10.1187/cbe.12-03-0026

Greiff, S. (2012). Assessment and theory in complex problem solving – A continuing contradiction? *Journal of Educational and Developmental Psychology*, *2*(1), 49–56. https://doi.org/10.5539/jedp.v2n1p49

Grube, C. (2010). Kompetenzen naturwissenschaftlicher Erkenntnisgewinnung (Doctoral dissertation). Retrieved from https://kobra.bibliothek.uni-kassel.de/bitstream/urn:nbn:de:hebis:34-2011041537247/3/DissertationChristianeGrube.pdf

Gustafsson, J.-E., & Åberg-Bengtsson, L. (2010). Unidimensionality and interpretability of psychological instruments. In S. E. Embretson (Ed.), *Measuring psychological constructs: Advances in model-based approaches.* (pp. 97–121). Washington, DC: American Psychological Association.

Haberman, S. J. (2007). When can subscores have value? *Journal of Educational and Behavioral Statistics*, *33*(2), 204–229. https://doi.org/10.3102/1076998607302636

Halpern, D. F., Millis, K., Graesser, A. C., Butler, H., Forsyth, C., & Cai, Z. (2012). Operation ARA: A computerized learning game that teaches critical thinking and scientific reasoning. *Thinking Skills and Creativity*, *7*(2), 93–100. https://doi.org/10.1016/j.tsc.2012.03.006

Hammann, M., Phan, T. H., & Bayrhuber, H. (2008). Experimentieren als Problemlösen: Lässt sich das SDDS-Modell nutzen, um unterschiedliche Dimensionen beim Experimentieren zu messen? In M. Prenzel, I. Gogolin, & H.-H. Krüger (Eds.), *Kompetenzdiagnostik* (pp. 33–49). Wiesbaden: VS Verlag für Sozialwissenschaften.

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

Harlen, W. (1999). Purposes and procedures for assessing science process skills. *Assessment in Education: Principles, Policy & Practice*, *6*(1), 129–144. https://doi.org/10.1080/09695949993044

Hartig, J., & Höhler, J. (2009). Multidimensional IRT models for the assessment of competencies. *Studies in Educational Evaluation*, *35*(2–3), 57–63. https://doi.org/10.1016/j.stueduc.2009.10.002

Hayduk, L. A. (1987). *Structural equation modeling with LISREL: Essentials and advances.* Baltimore, MD: Johns Hopkins University Press.

Heene, M. (2007). Konstruktion und Evaluation eines Studierendenauswahlverfahrens für Psychologie an der Universität Heidelberg (Doctoral dissertation). Retrieved from http://archiv.ub.uni-heidelberg.de/volltextserver/7727/1/Diss_Text7_final_published.pdf

Herring, J. P. (1918). Measurement of some abilities in scientific thinking. *Journal of Educational Psychology*, *9*(10), 535–558.

Hestenes, D., Wells, M., & Swackhamer, G. (1992). Force concept inventory. *The Physics Teacher*, *30*(3), 141–158.

Hickey, D. T., Ingram-Goble, A. A., & Jameson, E. M. (2009). Designing assessments and assessing designs in virtual educational environments. *Journal of Science Education and Technology*, *18*(2), 187–208. https://doi.org/10.1007/s10956-008-9143-1

Holbrook, J., & Rannikmae, M. (2009). The meaning of scientific literacy. *International Journal of Environmental and Science Education*, *4*(3), 275–288.

Huffman, D., & Heller, P. (1995). What does the force concept inventory actually measure? *The Physics Teacher*, *33*(3), 138. https://doi.org/10.1119/1.2344171

Inhelder, B., & Piaget, J. (1958). *The growth of logical thinking from childhood to adolescence: An essay on the construction of formal operational structure*. London: Routledge & Kegan Pau.

Karmiloff-Smith, A. (2012). Is development domain specific or domain general? A third alternative. In J. Shrager & S. Carver (Eds.), *The journey from child to scientist: Integrating cognitive development and the education sciences*. Washington, DC: American Psychological Association.

Khishfe, R. (2008). The development of seventh graders' views of nature of science. *Journal of Research in Science Teaching*, *45*(4), 470–496. https://doi.org/10.1002/tea.20230

Kind, P. M. (2013). Establishing assessment scales using a novel disciplinary rationale for scientific reasoning: Assessment scales with scientific reasoning. *Journal of Research in Science Teaching*, *50*(5), 530–560. https://doi.org/10.1002/tea.21086

Kind, P. M., & Osborne, J. (2016). Styles of scientific reasoning - a cultural rationale for science education? *Science Education*, *in press*. https://doi.org/doi:10.1002/sce.21251

Klahr, D., & Dunbar, K. (1988). Dual space search during scientific reasoning. *Cognitive Science*, *12*, 1–48.

Klos, S. (2009). *Kompetenzförderung im naturwissenschaftlichen Anfangsunterricht – Der Einfluss eines integrierten Unterrichtskonzepts*. Berlin: Logos Verlag.

Klos, S., Henke, C., Kieren, C., Walpuski, M., & Sumfleth, E. (2008). Naturwissenschaftliches Experimentieren und chemisches Fachwissen – Zwei verschiedene Kompetenzen. *Zeitschrift Für Pädagogik*, *54*(3), 304–321.

Koerber, S., Mayer, D., Osterhaus, C., Schwippert, K., & Sodian, B. (2015). The development of scientific thinking in elementary school: A comprehensive inventory. *Child Development*, *86*(1), 327–336. https://doi.org/10.1111/cdev.12298

Kok-Auntoh, & Woolnough, B. E. (1994). Science process skills: Are they generalisable? *Research in Science & Technological Education*, *12*(1), 31–42. https://doi.org/10.1080/0263514940120105

Kuhn, D. (2002). What is scientific thinking and how does it develop? In U. Goswami (Ed.), *Handbook of childhood cognitive development* (pp. 371–393). Oxford: Blackwell.

Kuhn, D., Iordanou, K., Pease, M., & Wirkala, C. (2008). Beyond control of variables: What needs to develop to achieve skilled scientific thinking? *Cognitive Development*, *23*(4), 435–451. https://doi.org/10.1016/j.cogdev.2008.09.006

Kuhn, D., Schauble, L., & Garcia-Mila, M. (1992). Cross-domain development of scientific reasoning. *Cognition and Instruction*, *9*(4), 285–327. https://doi.org/10.1207/s1532690xci0904_1

Lasry, N., Rosenfield, S., Dedic, H., Dahan, A., & Reshef, O. (2011). The puzzling reliability of the Force Concept Inventory. *American Journal of Physics*, *79*(9), 909–912. https://doi.org/10.1119/1.3602073

Lawson, A. E. (1978). The development and validation of a classroom test of formal reasoning. *Journal of Research in Science Teaching*, *15*(1), 11–24.

Lawson, A. E. (2000). Classroom Test of Scientific Reasoning, revised edition. Retrieved from http://modeling.asu.edu/modeling/weblinks.html

Lawson, A. E., Alkhoury, S., Benford, R., Clark, B. R., & Falconer, K. A. (2000). What kinds of scientific concepts exist. Concept construction and intellectual development in college biology. *Journal of Research in Science Teaching*, *37*(9), 996–1018.

Lawson, A. E., Clark, B., Cramer-Meldrum, E., Falconer, K. A., Sequist, J. M., & Kwon, Y.-J. (2000). Development of scientific reasoning in college biology: Do two levels of general hypothesis-testing skills exist? *Journal of Research in Science Teaching*, *37*(1), 81–101.

Lazonder, A. W., & Harmsen, R. (2016). Meta-analysis of inquiry-based learning: Effects of guidance. *Review of Educational Research*, *86*(3), 681–718. https://doi.org/10.3102/0034654315627366

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. https://doi.org/10.1002/tea.10034

Li, M., Ruiz-Primo, M. A., & Shavelson, R. J. (2006). Towards a science achievement framework: The case of TIMSS 1999. In S. J. Howie & T. Plomp (Eds.), *Contexts of learning mathematics and science: Lessons learned from TIMSS* (pp. 291–311). Routledge.

Liang, L. L., Chen, S., Chen, X., Kaya, O. N., Adams, A. D., Macklin, M., & Ebenezer, J. (2008). Assessing preservice elementary teachers' views on the nature of scientific knowledge: A dual-response instrument. *Asia-Pacific Forum on Science Learning and Teaching*, *9*(1), 1–20.

Linn, M. C., & Rice, M. (1979). A measure of scientific reasoning: The Springs task. *Journal of Educational Measurement*, *16*(1), 55–58.

Livermore, A. H. (1964). The process approach of the AAAS commission on science education. *Journal of Research in Science Teaching*, *2*(4), 271–282.

Macy, M. J., & Wood, H. B. (1951). Test of critical thinking. *University of Oregon Curriculum Bulletin*, *99*, 1–13.

Martin, M. O., & Mullis, I. V. S. (Eds.). (2012). *Methods and procedures in TIMSS and PIRLS 2011*. Chestnut Hill, MA: TIMSS & PIRLS International Study Center, Boston College.

Martin, M. O., Mullis, I. V. S., Foy, P., & Stanco, G. M. (2012). *TIMSS 2011 international results in science*. Chestnut Hill, MA: TIMSS & PIRLS International Study Center, Boston College.

Mayer, D. (2012). Die Modellierung des wissenschaftlichen Denkens im Grundschulalter (Doctoral dissertation). Retrieved from http://edoc.ub.uni-muenchen.de/14497/

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

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

Mayer, R. E., & Duemler, D. (1990). Empirical constraints on theories of scientific reasoning: Reply to Baron (1990). *Journal of Educational Psychology*, *82*(2), 393–395.

Mc Eneaney, E. H. (2003). The worldwide cachet of scientific literacy. *Comparative Education Review*, *47*(2), 217–237.

McDonald, R. P. (1965). Difficulty factors and non-linear factor analysis. *British Journal of Mathematical and Statistical Psychology*, *18*(1), 11–23.

Meerah, T. S. M., Osman, K., Zakaria, E., Ikhsan, Z. H., Krish, P., Lian, D. K. C., & Mahmod, D. (2012). Developing an instrument to measure research skills. *Procedia - Social and Behavioral Sciences*, *60*, 630–636. https://doi.org/10.1016/j.sbspro.2012.09.434

Millsap, R. E. (1995). Measurement invariance, predictive invariance, and the duality paradox. *Multivariate Behavioral Research*, *30*(4), 577–605. https://doi.org/10.1207/s15327906mbr3004_6

Millsap, R. E. (2007). Invariance in measurement and prediction revisited. *Psychometrika*, *72*(4), 461–473.

Molenaar, D., & Borsboom, D. (2013). The formalization of fairness: Issues in testing for measurement invariance using subtest scores. *Educational Research and Evaluation*, *19*(2–3), 223–244. https://doi.org/10.1080/13803611.2013.767628

Molitor, L. L., & George, K. D. (1976). Development of a test of science process skills. *Journal of Research in Science Teaching*, *13*(5), 405–412.

Moosbrugger, H., & Kelava, A. (2008). *Testtheorie und Fragebogenkonstruktion.* Berlin: Springer.

Morris, G. A., Harshman, N., Branum-Martin, L., Mazur, E., Mzoughi, T., & Baker, S. D. (2012). An item response curves analysis of the Force Concept Inventory. *American Journal of Physics*, *80*(9), 825–831. https://doi.org/10.1119/1.4731618

Muis, K. R., Bendixen, L. D., & Haerle, F. C. (2006). Domain-generality and domain-specificity in personal epistemology research: Philosophical and empirical reflections in the development of a theoretical framework. *Educational Psychology Review*, *18*(1), 3–54. https://doi.org/10.1007/s10648-006-9003-6

Mullis, I. V. S., Martin, M. O., Ruddock, G. J., O'Sullivan, C. Y., & Preuschoff, C. (2009). *TIMSS 2011 assessment frameworks*. Chestnut Hill, MA: TIMSS & PIRLS International Study Center, Boston College.

National Assessment Governing Board. (2007). *Science assessment and item specifications for the 2009 National Assessment of Educational Progress*. Washington, DC: National Assessment Governing Board.

National Research Council [NRC]. (1996). *National science education standards*. Washington, DC: National Academies Press.

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

Neumann, I., Neumann, K., & Nehm, R. (2011). Evaluating instrument quality in science education: Rasch-based analyses of a nature of science test. *International Journal of Science Education*, *33*(10), 1373–1405. https://doi.org/10.1080/09500693.2010.511297

Niaz, M. (1995). Enhancing thinking skills: Domain specific/domain general strategies - A dilemma for science education. *Instructional Science*, *22*(6), 413–422.

Noroozi, O., Weinberger, A., Biemans, H. J. A., Mulder, M., & Chizari, M. (2012). Argumentation-based computer supported collaborative learning (ABCSCL): A synthesis of 15 years of research. *Educational Research Review*, *7*(2), 79–106. https://doi.org/10.1016/j.edurev.2011.11.006

Norris, S. P., Phillips, L. M., & Burns, D. P. (2014). Conceptions of scientific literacy: Identifying and evaluating their programmatic elements. In M. R. Matthews (Ed.), *International handbook of research in history, philosophy and science teaching* (pp. 1317–1344). Dordrecht: Springer Netherlands.

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

Opitz, A., Heene, M., & Fischer, F. (2015). If scientific reasoning is what the tests test, then what is scientific reasoning? A review. Poster presented at the 96. Annual Meeting of the American Educational Research Association, Chicago, IL.

Organisation for Economic Co-operation and Development [OECD]. (2006). *Assessing scientific, reading and mathematical literacy: A framework for PISA 2006*. Paris: OECD.

Organisation for Economic Co-operation and Development [OECD]. (2007). *Science competencies for tomorrow's world. Volume I: Analysis.* Paris: OECD.

Organisation for Economic Co-operation and Development [OECD]. (2009). *PISA 2006 technical report*. Paris: OECD.

Osbeck, L. M. (2014). Scientific reasoning as sense-making: Implications for qualitative inquiry. *Qualitative Psychology*, *1*(1), 34–46. https://doi.org/10.1037/qup0000004

Osborne, J. (2010). Arguing to learn in science: The role of collaborative, critical discourse. *Science*, *328*(5977), 463–466. https://doi.org/10.1126/science.1183944

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., Collins, S., Ratcliffe, M., Millar, R., & Duschl, R. (2003). What "ideas-about-science" should be taught in school science? A Delphi study of the expert community. *Journal of Research in Science Teaching*, *40*(7), 692–720. https://doi.org/10.1002/tea.10105

Pant, H. A., Stanat, P., Schroeders, U., Roppelt, A., Siegle, T., & Pöhlmann, C. (Eds.). (2013). *IQB-Ländervergleich 2012: Mathematische und naturwissenschaftliche Kompetenzen am Ende der Sekundarstufe I.* Münster: Waxmann.

Pedaste, M., Mäeots, M., Siiman, L. A., de Jong, T., van Riesen, S. A. N., Kamp, E. T., … Tsourlidaki, E. (2015). Phases of inquiry-based learning: Definitions and the inquiry cycle. *Educational Research Review*, *14*, 47–61. https://doi.org/10.1016/j.edurev.2015.02.003

Pellegrino, J. W. (2013). Proficiency in science: Assessment challenges and opportunities. *Science*, *340*(6130), 320–323. https://doi.org/10.1126/science.1232065

Pellegrino, J. W., Wilson, M. R., Koenig, J. A., & Beatty, A. S. (Eds.). (2014). *Developing assessments for the next generation science standards*. Washington, DC: National Academies Press.

Perkins, D. N., & Salomon, G. (1989). Are cognitive skills context-bound? *Educational Researcher*, *18*(1), 16–25.

Piekny, J., & Maehler, C. (2013). Scientific reasoning in early and middle childhood: The development of domain-general evidence evaluation, experimentation, and hypothesis generation skills. *British Journal of Developmental Psychology*, *31*(2), 153–179. https://doi.org/10.1111/j.2044-835X.2012.02082.x

Planinic, M., Ivanjek, L., & Susac, A. (2010). Rasch model based analysis of the Force Concept Inventory. *Physical Review Special Topics - Physics Education Research*, *6*(1), 1–11. https://doi.org/10.1103/PhysRevSTPER.6.010103

Pratt, C., & Hacker, R. G. (1984). Is Lawson's Classroom Test of Formal Reasoning valid? *Educational and Psychological Measurement*, *44*(2), 441–448. https://doi.org/10.1177/0013164484442025

Preckel, F., & Thiemann, H. (2003). Online- versus paper-pencil-version of a high potential intelligence test. *Swiss Journal of Psychology*, *62*(2), 131–138.

Prudon, P. (2016). Testing predicted clusters: A new approach to this powerful research tool, illustrated through a questionnaire on obsessive-compulsive disorder. *Comprehensive Psychology*, *5*, 1–20. https://doi.org/10.1177/2165222816646237

Rasch, G. (1960). *Probabilistic models for some intelligence and achievement tests*. Copenhagen: Danish Institute for Educational Research.

Reise, S. P., Bonifay, W. E., & Haviland, M. G. (2013). Scoring and modeling psychological measures in the presence of multidimensionality. *Journal of Personality Assessment*, *95*(2), 129–140. https://doi.org/10.1080/00223891.2012.725437

Reise, S. P., Widaman, K. F., & Pugh, R. H. (1993). Confirmatory factor analysis and item response theory: Two approaches for exploring measurement invariance. *Psychological Bulletin*, *114*(3), 552–566.

Revelle, W. (2016). *Psych: Procedures for psychological, psychometric, and personality research* [R package]. Evanston, IL: Northwestern University. Retrieved from http://CRAN.R-project.org/package=psych

Revelle, W., & Zinbarg, R. E. (2009). Coefficients alpha, beta, omega, and the glb: Comments on Sijtsma. *Psychometrika*, *74*(1), 145–154.

Roberts, R., & Gott, R. (2004). A written test for procedural understanding: A way forward for assessment in the UK science curriculum? *Research in Science & Technological Education*, *22*(1), 5–21. https://doi.org/10.1080/0263514042000187511

Roskam, E. E. (1989). Operationalization, a superfluous concept. *Quality and Quantity*, *23*(3–4), 237–275.

Ross, J. A., & Maynes, F. J. (1983). Development of a test of experimental problem-solving skills. *Journal of Research in Science Teaching*, *20*(1), 63–75.

Rosseel, Y. (2012). Lavaan: An R package for structural equation modeling. *Journal of Statistical Software*, *48*(2), 1–36.

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

Rudolph, J. L., & Horibe, S. (2016). What do we mean by science education for civic engagement? *Journal of Research in Science Teaching*, *53*, 805–820. https://doi.org/doi: 10.1002/tea.21303

Ryder, J. (2001). Identifying science understanding for functional scientific literacy. *Studies in Science Education*, *36*(1), 1–44. https://doi.org/10.1080/03057260108560166

Sandoval, W. A. (2005). Understanding students' practical epistemologies and their influence on learning through inquiry. *Science Education*, *89*(4), 634–656. https://doi.org/10.1002/sce.20065

Schauberger, G. (2016). DIFlasso: A penalty approach to differential item functioning in rasch models (Version 1.0-2.) [R package]. Retrieved from https://CRAN.R-project.org/package=DIFlasso

Schauble, L. (1996). The development of scientific reasoning in knowledge-rich contexts. *Developmental Psychology*, *32*(1), 102–119.

Schizas, D., Psillos, D., & Stamou, G. (2016). Nature of science or nature of the sciences? *Science Education*, *100*(4), 706–733. https://doi.org/10.1002/sce.21216

Scholl, R., & Nickelsen, K. (2015). Discovery of causal mechanisms: Oxidative phosphorylation and the Calvin–Benson cycle. *History and Philosophy of the Life Sciences*, *37*(2), 180–209. https://doi.org/10.1007/s40656-015-0061-2

Schunn, C. D., & Anderson, J. R. (1999). The generality/specificity of expertise in scientific reasoning. *Cognitive Science*, *23*(3), 337–370.

Scott, T. F., & Schumayer, D. (2015). Students' proficiency scores within multitrait item response theory. *Physical Review Special Topics - Physics Education Research*, *11*(2), 1–12. https://doi.org/10.1103/PhysRevSTPER.11.020134

Scott, T. F., Schumayer, D., & Gray, A. R. (2012). Exploratory factor analysis of a Force Concept Inventory data set. *Physical Review Special Topics - Physics Education Research*, *8*(2), 1–10. https://doi.org/10.1103/PhysRevSTPER.8.020105

Shavelson, R. J., Baxter, G. P., & Pine, J. (1991). Performance assessment in science. *Applied Measurement in Education*, *4*(4), 347–362.

Shavelson, R. J., & Huang, L. (2003). Responding responsibly to the frenzy to assess learning in higher education. *Change: The Magazine of Higher Learning*, *35*(1), 11–19.

Shaw, T. J. (1983). The effect of a process-oriented science curriculum upon problem-solving ability. *Science Education*, *67*(5), 615–623.

Shute, V. J., & Kim, Y. J. (2013). Formative and stealth assessment. In J. M. Spector, M. D. Merrill, J. Elen, & Bishop (Eds.), *Handbook of research on educational communications and technology* (4th ed., pp. 311–323). New York, NY: Lawrence Erlbaum Associates.

Simon, H. (1966). Scientific discovery and psychology of problem solving. In R. Colony (Ed.), *Mind and cosmos* (pp. 22–40). Pittsburgh: Editions Pittsburgh University Press.

Sinatra, G. M., & Chinn, C. A. (2012). Thinking and reasoning in science: Promoting epistemic conceptual change. In K. R. Harris, S. Graham, T. Urdan, A. G. Bus, S. Major, & H. L. Swanson (Eds.), *APA educational psychology handbook, vol. 3: Application to learning and teaching.* (pp. 257–282). Washington, DC: American Psychological Association.

Sirum, K., & Humburg, J. (2011). The Experimental Design Ability Test (EDAT). *Bioscene*, *37*(1), 8–16.

Soares, T. M., Goncalves, F. B., & Gamerman, D. (2009). An integrated Bayesian model for DIF analysis. *Journal of Educational and Behavioral Statistics*, *34*(3), 348–377. https://doi.org/10.3102/1076998609332752

Solomon, M. (2001). *Social empiricism*. Cambridge, MA: MIT press.

Soobard, R., & Rannikmäe, M. (2011). Assessing student's level of scientific literacy using interdisciplinary scenarios. *Science Education International*, *22*(2), 133–144.

Spearman, C. (1927). *The abilities of man.* London: Macmillan.

State Council of Higher Education for Virginia. (2007). *Guidelines for assessment of student learning*. Retrieved from http://www.schev.edu/docs/default-source/Reports-and-Studies/2007/2007assessmentguidelines.pdf?sfvrsn=4

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*(5), 721–732. https://doi.org/10.1080/02602938.2016.1164830

Strobel, C., Heitzmann, N., Strijbos, J.-W., Kollar, I., & Fischer, M. (2016). Watching people fail: Improving diagnostic competence by providing peer feedback on erroneous diagnoses. Paper presented at the 97. Annual Meeting of the American Educational Research Association, Washington, DC.

Strobl, C., Kopf, J., & Zeileis, A. (2015). Rasch trees: A new method for detecting differential item functioning in the Rasch model. *Psychometrika*, *80*(2), 289–316.

Strobl, C., Malley, J., & Tutz, G. (2009). An introduction to recursive partitioning: Rationale, application, and characteristics of classification and regression trees, bagging, and random forests. *Psychological Methods*, *14*(4), 323–348. https://doi.org/10.1037/a0016973

Su, J.-M., Lin, H.-Y., Tseng, S.-S., & Lu, C.-J. (2011). OPASS: An online portfolio assessment and diagnosis scheme to support web-based scientific inquiry experiments. *Turkish Online Journal of Educational Technology*, *10*(2), 151–173.

Sundre, D. L. (2008). *The Scientific Reasoning Test, Version 9 (SR-9) test manual*. Harrisonburg, VA: Center for Assessment and Research Studies.

Szu, E., & Osborne, J. (2011). A Bayesian approach to scientific reasoning: Implications for science education. In M. S. Khine (Ed.), *Perspectives on scientific argumentation* (pp. 55–71). Dordrecht: Springer.

Tamir, P., Nussinovitz, R., & Friedler, Y. (1982). The design and use of a practical tests assessment inventory. *Journal of Biological Education*, *16*(1), 42–50. https://doi.org/10.1080/00219266.1982.9654417

The Royal Society. (2014). *Vision for science and mathematics education*. London: Royal Society.

Thelk, A. D., & Hoole, E. R. (2006). What are you thinking? Postsecondary student think-alouds of scientific and quantitative reasoning items. *The Journal of General Education*, *55*(1), 17–39.

Tibshirani, R. (1996). Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society. Series B (Methodological)*, *58*(1), 267–288.

Timmerman, B. E. C., Strickland, D. C., Johnson, R. L., & Payne, J. R. (2011). Development of a "universal" rubric for assessing undergraduates' scientific reasoning skills using scientific writing. *Assessment & Evaluation in Higher Education*, *36*(5), 509–547. https://doi.org/10.1080/02602930903540991

Tobin, K. G., & Capie, W. (1981). The development and validation of a group test of logical thinking. *Educational and Psychological Measurement*, *41*(2), 413–423. https://doi.org/10.1177/001316448104100220

Tricot, A., & Sweller, J. (2014). Domain-specific knowledge and why teaching generic skills does not work. *Educational Psychology Review*, *26*(2), 265–283. https://doi.org/10.1007/s10648-013-9243-1

Tucker, B. (2009). The next generation of testing. *Educational Leadership*, *67*(3), 48–53.

Turkan, S., & Liu, O. L. (2012). Differential performance by English language learners on an inquiry-based science assessment. *International Journal of Science Education*, *34*(15), 2343–2369. https://doi.org/10.1080/09500693.2012.705046

Tutz, G., & Berger, M. (2016). Item-focussed trees for the identification of items in differential item functioning. *Psychometrika*, *81*(3), 727–750. https://doi.org/10.1007/s11336-015-9488-3

Tutz, G., & Schauberger, G. (2015). A penalty approach to differential item functioning in Rasch models. *Psychometrika*, *80*(1), 21–43.

United States National Assessment Governing Board, WestEd (Organization), & Council of Chief State School Officers. (2010). *Science framework for the 2011 National Assessment of Educational Progress*. Washington, DC: National Assessment Governing Board, US Dept. of Education.

U.S. Department of Education, Institute of Education Sciences, National Center for Education Statistics, National Assessment of Educational Progress [NAEP]. (2000). *2000 Science assessment*.

U.S. Department of Education, Institute of Education Sciences, National Center for Education Statistics, National Assessment of Educational Progress [NAEP]. (2005). *2005 Science assessment*.

U.S. Department of Education, Institute of Education Sciences, National Center for Education Statistics, National Assessment of Educational Progress [NAEP]. (2009). *2009 Sciene assessment*.

van der Maas, H. L. J., Dolan, C. V., Grasman, R. P. P. P., Wicherts, J. M., Huizenga, H. M., & Raijmakers, M. E. J. (2006). A dynamical model of general intelligence: The

positive manifold of intelligence by mutualism. *Psychological Review*, *113*(4), 842–861. https://doi.org/10.1037/0033-295X.113.4.842

van Gigch, J. P. (2002). Comparing the epistemologies of scientific disciplines in two distinct domains: Modern physics versus social sciences. II: Epistemology and knowledge characteristics of the "new" social sciences. *Systems Research and Behavioral Science*, *19*(6), 551–562. https://doi.org/10.1002/sres.466

Vleeschouwer, M., Schubart, C. D., Henquet, C., Myin-Germeys, I., van Gastel, W. A., Hillegers, M. H., … Derks, E. M. (2014). Does assessment type matter? A measurement invariance analysis of online and paper and pencil assessment of the Community Assessment of Psychic Experiences (CAPE). *PloS One*, *9*(1), 1–9.

Ward, W. C., Frederiksen, N., & Carlson, S. B. (1980). Construct validity of free-response and machine-scorable forms of a test. *Journal of Educational Measurement*, *17*(1), 11–29.

Wecker, C., Hetmanek, A., & Fischer, F. (2016). Zwei Fliegen mit einer Klappe? Fachwissen und fächerübergreifende Kompetenzen gemeinsam fördern. *Unterrichtswissenschaft*, *44*(3), 226–238.

Weinberger, A., Stegmann, K., & Fischer, F. (2010). Learning to argue online: Scripted groups surpass individuals (unscripted groups do not). *Computers in Human Behavior*, *26*(4), 506–515. https://doi.org/10.1016/j.chb.2009.08.007

Weld, J., Stier, M., & McNew-Birren, J. (2011). The development of a novel measure of scientific reasoning growth among college freshmen: The Constructive Inquiry Science Reasoning Skills Test. *Journal of College Science Teaching*, *40*(4), 101–107.

Wellnitz, N., Fischer, H. E., Kauertz, A., Mayer, J., Neumann, I., Pant, H. A., … Walpuski, M. (2012). Evaluation der Bildungsstandards – Eine fächerübergreifende

Testkonzeption für den Kompetenzbereich Erkenntnisgewinnung. *Zeitschrift Für Didaktik Der Naturwissenschaften*, *18*, 261–291.

White, B., Stains, M., Escriu-Sune, M., Medaglia, E., Rostamnjad, L., Chinn, C., & Sevian, H. (2011). A novel instrument for assessing students' critical thinking abilities. *Journal of College Science Teaching*, *40*(5), 102–107.

Wiliam, D. (2001). What is wrong with our educational assessment and what can be done about it? *Education Review*, *15*(1), 57–62.

Wiliam, D. (2010). What counts as evidence of educational achievement? The role of constructs in the pursuit of equity in assessment. *Review of Research in Education*, *34*(1), 254–284. https://doi.org/10.3102/0091732X09351544

Wiliam, D. (2011). What is assessment for learning? *Studies in Educational Evaluation*, *37*(1), 3–14. https://doi.org/10.1016/j.stueduc.2011.03.001

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

Wu, M., Donovan, J., Hutton, P., & Lennon, M. (2008). *National Assessment Program--science literacy year 6 technical report, 2006* (p. 134). Carlton, South Victoria: Ministerial Council on Education, Employment, Training and Youth Affairs.

Xu, F., & Kushnir, T. (2013). Infants are rational constructivist learners. *Current Directions in Psychological Science*, *22*(1), 28–32. https://doi.org/10.1177/0963721412469396

Ziegler, M., & Hagemann, D. (2015). Testing the unidimensionality of items: Pitfalls and loopholes. *European Journal of Psychological Assessment*, *31*(4), 231–237. https://doi.org/10.1027/1015-5759/a000309

Zimmerman, C. (2000). The development of scientific reasoning skills. *Developmental Review*, *20*(1), 99–149. https://doi.org/10.1006/drev.1999.0497

Zimmerman, C. (2007). The development of scientific thinking skills in elementary and middle school. *Developmental Review*, *27*(2), 172–223. https://doi.org/10.1016/j.dr.2006.12.001

Zlatkin-Troitschanskaia, O., Shavelson, R. J., & Kuhn, C. (2015). The international state of research on measurement of competency in higher education. *Studies in Higher Education*, *40*(3), 393–411. https://doi.org/10.1080/03075079.2015.1004241

Zumbo, B. (1999). *A handbook on the theory and methods of differential item functioning (DIF).* Ottawa: National Defense Headquarters.

Zyve, D. (1939). *Stanford scientific aptitude test for high school and college students*. Palo Alto, CA: Stanford University Press.

Appendix

Appendix A: Coding scheme for the sorting of scientific reasoning skill descriptions (Study 1)

Appendix B: Coding scheme for rating the domain dependency of the CTSR items (Study 2)

Appendix A: Coding scheme for the sorting of scientific reasoning skill descriptions (Study 1)

General remarks

- Every description should only receive one code. It should be the code of the skill that fits the description best.
- If one description absolutely covers two different skills (maybe because they are connected with "and" in the author's description) it is possible to give a second code; in that case always use the lower number as the first code and the higher number as the second code.
- The basis for this coding scheme is the description of the eight epistemic activities by F. Fischer et al. (2014).
- Whenever the word "research" is used it does not only relate to research conducted by professional scientists.
- Sometimes the context of the other descriptions in a test can give a hint about the classification.
- The descriptions named under "*Example(s)*" are real descriptions used by test authors; *clarifications* are more general examples that elaborate on a sorting criteria; they are meant to help to understand the sorting criterion better, not necessarily to code a skill.

Overview of codes

Code	Scientific reasoning skill
1	Problem identification
2	Questioning
3	Hypothesis generation
4	Construction and redesign of artefacts
5	Evidence generation
6	Evidence evaluation
7	Drawing conclusions
8	Communicating and scrutinising
9	Other skill / not fitting into the other categories

Detailed description with examples

(1) Problem identification

- Recognizing a problem that needs to be investigated
- Building a problem representation
- Exception: if a description addresses shortcomings in somebody else's work, it is sorted into *communicating and scrutinising*

Example(s):

- “Formulating problems”

(2) Questioning

- Formulating a question for research

Example(s):

- “Presenting questions”

(3) Hypothesis generation

- Formulating hypotheses
- Making predictions for an experiment (vs. on the basis of evidence -> *drawing conclusions*); *clarification*: what is meant are statements like “one group will be better than the other group”
- Recognizing the hypothesis in a study
- Judging the scientific quality of a hypothesis; *clarification:* what is meant by this are characteristics of a hypothesis like “is it possible to falsify it” or “is there a way to test this empirically”

Example(s):

- “Formulating and Judging Ideas / Hypotheses”

(4) Construction and redesign of artefacts

- Tasks that ask for the development or redesign of an artefact; not to be confused with the construction of an experimental design, which is sorted into *evidence generation*; *clarification:* an artefact would be a prototype developed by an engineer or a learning intervention developed by a teacher student

Example(s):

- [No examples could be found]

(5) Evidence generation

- Descriptions about the control of variables strategy

- Descriptions connected with identifying variables (independent variable, dependent variable, covariate…) in a study
- Descriptions connected with collecting evidence (in whichever way: observations, experiments…) and conducting an investigation
- This code also includes the understanding of how to construct a good investigation and which parts are needed for that; *clarification*: what is meant are things like "is there an experimental and a control group?", "were confounding variables controlled?", "was a variable manipulated?", "is there more than one point of measurement?", etc.

Example(s):

- "Describing and controlling variables"
- "Understanding that experiments need to be repeated"
- "Understanding the role of the control in experiment (identifying the implemented controls in an experiment)"
- "Testing hypotheses" [the context of this description makes it clear that it belongs into this category]
- "Describing observations (regarding their accuracy and completeness)" [see the last remark for category 8 for why this is sorted here]

(6) Evidence evaluation

- Descriptions connected with the analysis of data
- Understanding correlations, graphs and tables
- Interpreting that is close to the data / results (what does it mean in regard to the hypothesis?); *clarification:* a statement like "as stated in the hypothesis the means of men and women differ"
- Judgements about the quality of data (being the basis for a conclusion)

Example(s):

- "Lines of best fit"
- "Understanding the need of statistical data analyses"
- "classifying"
- "Data and Hypotheses (distinguish between correct restatement of data and valid hypotheses as the reasons for data presented in a graph)"
- "Supporting Data (determine if data from various experiments support one or another hypothesis)"
- "Verification (ability to recognize that observation which is necessary and sufficient to establish the validity of an inference)"

(7) Drawing conclusions

- Descriptions about conclusions (reconsidering an initial claim); also includes interpretations that are directly linked to conclusions; *clarification:* a statement like "women outperform men in this task"
- Decisions based on the evidence evaluation (what do I have to do now?); *clarification:* for example, a change in a theory or choosing a treatment
- Generalizations / what generalizations are possible
- Limitations of a study
- Making a prediction on the basis of evidence

Example(s):

- "Interpretation of data where some conclusions cannot be justified (selecting which conclusions are justified and which aren't)"
- "Conjecturing other interpretations of particular studies"
- "Warranting claims"
- "Identifying appropriate descriptions, explanations, and predictions"
- "Interpretation (weigh evidence and decide if generalizations or conclusions based upon the data given in tables and graphs are warranted)"
- "Inferring"

(8) Communicating and scrutinising

- Descriptions connected with the act of communicating results, also including subskills necessary for that (e.g. making correct citations is necessary to write a scientific paper)
- Descriptions connected with criticizing the work of others; *clarification*: for example, criticizing the assumptions of someone
- Inferences
- Exception: if it is a judgment about the quality of one specific other scientific practice, it gets sorted into that practice (e.g. the elements of a good hypothesis mentioned under (3) or the elements of a good experiment mentioned under (5))

Example(s):

- "Critiquing the trustworthiness of evidence and claims made by others"
- "(Correct use of) Graphs"
- "Create graphical representations of data"

(9) Other skill / not fitting into the other categories

- Mainly mathematical skills (reasoning about probabilities, proportions etc.)
- Specific skills not clearly related to a scientific investigation
- Content knowledge and the application of content knowledge

- Things related to the understanding of the nature of science; however: if it is related to the understanding of a specific scientific practice, it gets sorted into this practice
- Things related to the role of science in society
- Descriptions too vague and/or short to sort them anywhere else

Example(s):

- “Understanding the connection of concepts”
- “Application of knowledge to problems”
- “Assessing the logic and accuracy of statements related to chemistry”
- “Variation in measurements of the same quantity (explaining why variations in results exist)”
- “Understanding that it is not possible to prove a hypothesis”
- “Graph types and variables”
- “Identifying keywords to search for scientific information”
- “Giving reasons for the relevance of a study”
- “Identify a valid scientific argument”

Appendix B: Coding scheme for rating the domain dependency of the CTSR items (Study 2)

Studie: Evaluation eines Tests zum wissenschaftlichen Denken

Kodierschema zur Kontextabhängigkeit der Items des Lawson-Tests

Allgemeine Anmerkung: Als „Domänen" werden hier Physik bzw. Biologie verstanden

Die Kodierung besteht aus 2 Stufen, wobei die 2. Stufe oft nicht nötig sein wird.

Stufe 1: Alle Items sollen in einem ersten Schritt einer der folgenden Kategorien zugeordnet werden. Es ist dabei durchaus möglich, dass die Häufigkeit der einzelnen Kategorien ungleich verteilt ist und mache Kategorien gar nicht vergeben werden.

Kategorien:

Kat.	Beschreibung
0	Das Item besitzt keinen Kontext aus der Domäne des Beurteilers. Das Item kann also entweder überhaupt keiner Domäne zuzuordnen sein oder einer Domäne, die nicht der des Beurteilers entspricht.
1	Das Item ist in einem domänenspezifischen Kontext angesiedelt, aber domänenspezifische Anteile (siehe Stufe 2 unten) sind für die Lösung des Items nicht relevant.
2	Das Item ist in einem domänenspezifischen Kontext angesiedelt und hat domänenspezifische Anteile, deren Beherrschung zur Lösung hilfreich ist, allerdings können oder müssen auch nicht-domänenspezifische Anteile beherrscht werden, um die Aufgabe zu lösen. Unter diese Kategorie würde es also z.B. fallen, wenn mit domänenspezifischem Wissen ein Teil der Antwortmöglichkeiten ausgeschlossen werden können oder wenn die Bekanntheit eines bestimmten experimentellen Settings bei der Beantwortung hilft aber das Item auch ohne Kenntnis des Settings gelöst werden kann.
3	Das Item ist in einem domänenspezifischen Kontext angesiedelt und hat domänenspezifische Anteile, deren Beherrschung zur korrekten Beantwortung des Items zwingend notwendig ist.

Stufe 2: Bei Items, die in Kategorie 2 oder 3 fallen, soll dann noch eine zusätzliche Beschreibung erfolgen, worin (Ihrer Meinung nach) der domänenspezifische Anteil besteht. Dabei können wahrscheinlich v.a. die folgenden 3 Aspekte relevant sein:

- Kenntnis von domänenspezifischem Wissen
- Kenntnis domänenspezifischer Methodik
- Bekanntheit der domänenspezifischen Begriffe oder Phänomene (z.B. Fachterminologie oder ein bestimmter Experimentalaufbau)

Sollte es darüber hinaus / stattdessen andere domänenspezifische Anteile geben, die (Ihrer Meinung nach) für die Lösung des Items relevant sind, dürfen natürlich auch diese genannt werden!

Beispiele (vereinfacht) zur Veranschaulichung (größtenteils an Hand des domänenspezifischen Anteils „Wissen“):

Kat.	Beispiel	Erklärung
	Beispiel ohne Kontext	
0	Was ist 2 x 2?	Das Item ist keinem Kontext (bzw. der Mathematik) zuzuordnen
	Beispiele aus der Biologie	
1	2 Kühe haben jeweils 2 Kälber. Wie viele Tiere sind dies insgesamt?	Das Item besitzt zwar einen Kontext, den man der Biologie zuordnen kann, allerdings hilft biologisches Wissen nicht bei der Beantwortung der Frage
2	Es gibt dominante und rezessive Allele. Werden 2 Allele kombiniert, wie viele unterschiedliche Kombinationsmöglichkeiten gibt es (die Reihenfolge der Allele in einer Kombination spielt keine Rolle)?	Domänenspezifisches Wissen kann bei der Beantwortung dieser Frage helfen (z.B. wenn man einen Teil oder alle möglichen Kombinationen von 2 Allelen auswendig weiß), aber das Item kann genauso gelöst werden mit kombinatorischen Fähigkeiten. (Zudem könnte hier eventuell noch die Bekanntheit der Fachbegriffe eine Rolle spielen)
3	Wie kann es sein, dass eine dominant-rezessiv vererbte Eigenschaft in einer Generation vorliegt, ohne dass die Eigenschaft in der Elterngeneration auftrat?	Diese Frage kann nur mit domänenspezifischem Wissen über Erbgänge korrekt beantwortet werden.
	Beispiele aus der Physik	
1	Als Galileo Experimente auf der schiefen Ebene durchführte, variierte er das Material der Kugel und hielt andere Variablen konstant. Erklären Sie warum es im Allgemeinen bei Experimenten wichtig ist nur eine Variable zu verändern!	Die Aufgabe ist zwar durch den ersten Satz in einen physikalischen Kontext eingebettet aber zur korrekten Beantwortung der Frage ist es letztlich entscheidend, ob man situationsunabhängig die Variablenkontrollstrategie verstanden hat.
2	Die Formel zur Äquivalenz von Masse und Energie besagt, dass Energie (E) dem Produkt aus Masse (m) und Lichtgeschwindigkeit (c) zum Quadrat entspricht. Wie lautet die Formel?	Diese Frage kann mit domänenspezifischem Wissen beantwortet werden (z.B. wenn man die Formel auswendig weiß) aber man kann die korrekte Antwort auch genauso aus den Informationen in der Frage herleiten (mit mathematischem Wissen, das nicht spezifisch für die Physik ist).
3	Wie lautet die Formel zur Äquivalenz von Masse und Energie?	Im Gegensatz zur vorhergehenden Frage, kann diese Frage ausschließlich mit domänenspezifischem Wissen beantwortet werden.

Anmerkung: Hier sind wie gesagt nur Beispiele aufgeführt, bei denen (größtenteils) Wissen eine Rolle spielt. Es wäre aber z.B. auch denkbar, dass ein typischer Experimentalaufbau aus der Physik gezeigt wird zusammen mit Ergebnissen des Experiments und man muss eine Schlussfolgerung aus den Ergebnissen ziehen. Ein Physiker könnte dann dadurch einen Vorteil haben, dass der Experimentalaufbau schneller verstanden wird und man sich daher mehr auf die Ergebnisse und Schlussfolgerung konzentrieren kann.

Kodierbogen zur Kontextabhängigkeit der Items des Lawson-Tests

Name der / des Kodierenden:

Kodierung:

Item(paar)	Fragen	Kategorie	Domänenspezifischer Anteil (falls Kategorie 2 oder 3 gewählt wurde)
1	1 + 2		
2	3 + 4		
3	5 + 6		
4	7 + 8		
5	9 + 10		
6	11 + 12		
7	13 + 14		
8	15 + 16		
9	17 + 18		
10	19 + 20		
11	21 + 22		
12	23		
13	24		

Anmerkung: Im Lawson-Test ergeben (fast) immer 2 aufeinander folgende Fragen ein Item(paar). Die einzige Ausnahme bilden die letzten beiden Fragen, die kein Paar bilden, sondern einzelne Items darstellen. Daher ergeben sich aus 24 Fragen 13 Items.