Routledge
Taylor & Francis Group

Educational Research and Evaluation

An International Journal on Theory and Practice

ISSN: 1380-3611 (Print) 1744-4187 (Online) Journal homepage: https://www.tandfonline.com/loi/nere20

Measuring scientific reasoning – a review of test instruments

Ansgar Opitz, Moritz Heene & Frank Fischer

To cite this article: Ansgar Opitz, Moritz Heene & Frank Fischer (2017) Measuring scientific reasoning – a review of test instruments, Educational Research and Evaluation, 23:3-4, 78-101, DOI: 10.1080/13803611.2017.1338586

To link to this article: https://doi.org/10.1080/13803611.2017.1338586

EDUCATIONAL RESEARCH AND EVALUATION, 2017
VOL. 23, NOS. 3–4, 78–101
https://doi.org/10.1080/13803611.2017.1338586

Routledge
Taylor & Francis Group

Measuring scientific reasoning – a review of test instruments

Ansgar Opitz [a,b], Moritz Heene[a,b] and Frank Fischer[a,b]

[a]Department of Psychology, Ludwig-Maximilians-University Munich, Munich, Germany; [b]Munich Center of the Learning Sciences, Ludwig-Maximilians-University Munich, Munich, Germany

ABSTRACT

Education systems increasingly emphasize the importance of scientific reasoning skills such as *generating hypotheses* and *evaluating evidence*. Despite this importance, we do not know which tests of scientific reasoning exist, which skills they emphasize, how they conceptualize scientific reasoning, and how well they are evaluated. Therefore, this article reviews 38 scientific reasoning tests. They used to primarily consist of multiple-choice questions, but since then have become more diverse, even including tests that automatically analyse virtual experiments. Furthermore, this review revealed that the tests focus on the skills of *generating hypotheses, generating evidence, evaluating evidence, and drawing conclusions*. Additionally, conceptualizations of scientific reasoning have moved towards a domain-specific set of different but coordinated skills over the years. Finally, concluding from gaps in test evaluation, a future focus should be on testing theoretical assumptions, comparing different scientific reasoning tests, and how relevant test results are in predicting criterion variables like academic performance.

ARTICLE HISTORY
Received 1 August 2016
Accepted 30 May 2017

KEYWORDS
Scientific reasoning; test instrument; review; conceptualization; test format

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; National Research Council [NRC], 2012; Organisation for Economic Co-operation and Development [OECD], 2006). According to a recent meta-analysis on the effects of guidance on inquiry learning by Lazonder and Harmsen (2016), these scientific concepts and methods are seen as necessary to engage in inquiry learning and as the main target for guidance. They are also considered to be 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[1] 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, Phillips, & Burns, 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 the Programme

CONTACT Ansgar Opitz ansgar.opitz@psy.lmu.de

for International Student Assessment (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 1. We summarize these skills under the term *scientific reasoning*.

For the purpose of this review, we regard scientific reasoning as being different from (collaborative) scientific argumentation, which is a complex process of its own (Berland & McNeill, 2010). Engaging in discussions is a crucial part of science education (Osborne, 2010), and frameworks exist for how to assess these discussions (Clark & Sampson, 2008). However, a learner also has to acquire individual scientific reasoning skills. While engaging in scientific discussions is part of many scientific reasoning conceptualizations, the discussions 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). Of course, we need to know more both about assessments of (collaborative) scientific argumentation including their processes and social contexts and about (individual) scientific reasoning skills. As a first step, this work focuses on assessments that aim to measure scientific reasoning skills used by individuals outside of (collaborative) scientific argumentation situations. This review will also exclude tests related to the *nature of science* construct (Lederman, Abd-El-Khalick, Bell, & Schwartz, 2002) as it is focusing on knowledge about science. Therefore, nature of science is more about understanding scientific reasoning skills from an epistemic perspective and not about using them.

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 in 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 influential 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

Table 1. List of scientific reasoning skills used in this review (based on Fischer et al., 2014).

Skill name	Skill description
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.
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	Analysing 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.

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 has been suggested 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 (Fischer et al., 2014; NRC, 2012). Including more skills was a reaction to the criticism that older theories perpetuated the idea of a single "scientific method" that only recognizes controlled experiments as a way to gain knowledge but excludes other important aspects of scientific reasoning and the related skills (Bauer, 1994; NRC, 2012). In comparison to the SDDS model, the single skills were now no longer seen as the parts of a fixed process that always occurs in the same one-way sequence in a scientific endeavour. Instead, current theories allow the back-and-forth jumping between skills and the simultaneous engagement in several skills. This is seen as a closer representation of the actual work of scientists (NRC, 2012). In summary, 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.

Research goals

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 existing tests that we present in this review.

Research questions

(1) 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?
(2) 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.
(3) 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.
(4) 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 programme, 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, for example, 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.
(5) 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. Additionally, a meta-analysis on interventions targeting the control-of-variables strategy showed that higher effect sizes can be

found with real performance and open-ended tests compared to multiple-choice and virtual performance tests (Schwichow, Croker, Zimmerman, Höffler, & Härtig, 2016). It thus seems informative to explore the test formats of existing scientific reasoning tests.

(6) With all of these questions, we analysed if we can observe any trends over time.

Methods

Literature search

We conducted a literature search using the databases ERIC, PsycINFO, and PSYNDEX, as well as the Buros test review repository (http://buros.org/). 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 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 Fischer et al. (2014). Its eight skills are mentioned in Table 1. 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 Fischer et al. (2014). Another example is that both the conceptualizations by 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 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, 38 of the 84 sources were included and 46 were excluded. The following were the reasons to exclude a test: Several tests were excluded because none of their parts measured scientific reasoning but instead only measured one of the constructs that we defined as being different from scientific reasoning in the Introduction, for example, science knowledge or nature of science. Other reasons to exclude a test were that 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.

Test analysis

The selected tests were analysed regarding their year of development, target group(s), addressed skills, theoretical background(s) including the dimensionality of their structure, domain generality versus specificity assumptions, certain psychometric properties (reliability, content validity, construct validity – also including concurrent 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 has already been 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 Fischer et al. (2014). 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 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 analysed 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. It should be noted that "independent" does not imply that these skills are only thought of in isolation. Instead, it just means that the skills are not placed in a fixed process that always occurs in the same way and the same order. The determination of the domain generality versus specificity aspect was based on claims made by the test authors 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 versus 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 concurrent 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. To improve the assessment of the psychometric property checks, we not only included psychometric property checks from the original test articles but also searched for other articles validating the tests. These articles are also included in the overview of test properties in Table 2.

Results

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 2. 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.

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, 218 could be related to the eight main scientific reasoning skill categories and 40 had to be 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"

Table 2. 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, 2006)	OP	EG, EE, DC, CS, OT	S	s	B, C, P	R, CS, D/C, 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; Lawson, Alkhoury, Benford, Clark, & Falconer, 2000a; Lawson, Clark, et al., 2000b)	MC	EG. EE, OT	S, U	g	na	R, CT, CS, D/C, CR	+
Competence Scale for Learning Science	(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	–
Detector – Inquiry Intelligent Tutoring System	(Gobert, Sao Pedro, Raziuddin, & Baker, 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, 2008a; Hammann, Phan, Ehmer, & Grimm, 2008b)	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	–
Institut zur Qualitätsentwicklung im Bildungswesen (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, WestEd (Organization), & Council of Chief State School Officers, 2010)	MI	EG, EE, DC, OT	E, S	s	B, ES, P	na	+

(*Continued*)

Table 2. Continued.

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, 2008a; Donovan, Lennon, O'Connor, & Morrissey, 2008b; 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	(Feldon, Maher, Hurst, & Timmerman, 2015; Gilmore, Vieyra, Timmerman, Feldon, & Maher, 2015; Timmerman, Feldon, Maher, Strickland, & Gilmore, 2013; Timmerman, Strickland, Johnson, & Payne, 2011)	SC	HG, EG, EE, DC, CS, OT	U	n/u	B	CT, CR	–
Science Process Skill Test (SPST)	(Feyzioglu, Demirdag, Akyildiz, & Altun, 2012)	MC	HG, EG, EE, DC	S	n/u	C	R, CT, CS, CR	–
Science-P	(Koerber, Mayer, Osterhaus, Schwippert, & Sodian, 2015; Mayer, 2012; 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, Pulos, & Gans, 1981; Linn & Rice, 1979; Linn & Swiney, 1981)	OT	EG, CS	E, S, U	g	na	R, D/C, CR	–
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; Baird & Borich, 1987; Baird, Shaw, & McLarty, 1996; Burns, Okey, & Wise, 1985; Dillashaw & Okey, 1980; Padilla, Okey, & Dillashaw, 1983)	MC	HG, EG, EE	S	g	na	R, CT, CS, D/C, CR	–
Test Of Logical Thinking (TOLT)	(Tobin & Capie, 1981, 1982)	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 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" behavioural 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 concurrent validity), "D/C" divergent and/or concurrent validity, "CR" criterion validity; [g]test norms (criterion or population based): "+" norms are reported; "–" no norms reported.

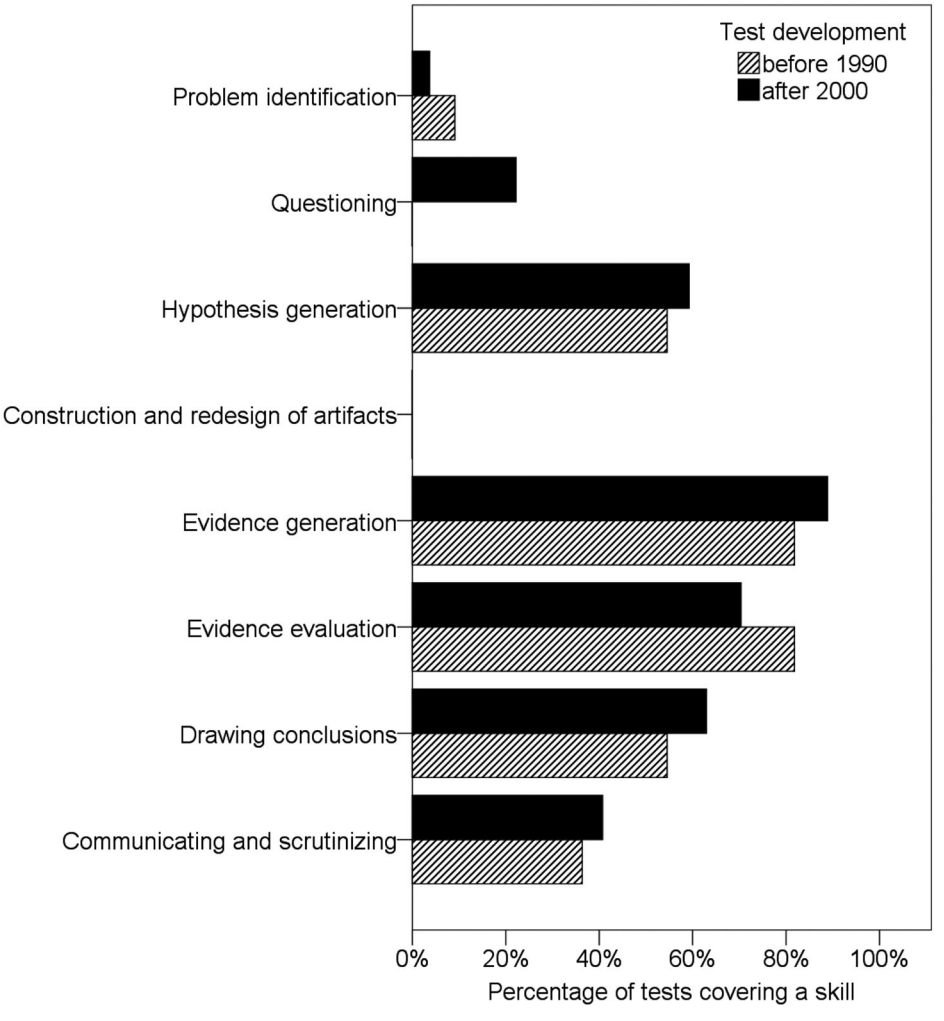

Figure 1. Scientific reasoning skills covered by the tests sorted by year of development.

category were referring to quantitative skills (9 descriptions) or to societal or ethical issues of science (5 descriptions).

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 the test authors referred to two separate theories as a basis for their test). The first category of theories – theories assuming 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 twice and once, respectively. Thus, taking into account 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 possess a multidimensional structure. This multidimensional structure most commonly takes the form of a problem-solving activity in which several skills have to be orchestrated.

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 dimension 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 principle, results from model tests could help to answer the question of dimensionality. However, results from model tests were few in number and differed widely. Looking at the results from various factor analyses, which were conducted with tests cited in this review, it is possible to find one-factor models (Germann, 1989; Gormally et al., 2012; Klos et al., 2008; Roberts & Gott, 2004; Tobin & Capie, 1981). The names of these factors align with the construct the overall test is supposed to measure. For instance, they are described as a single scientific literacy (Gormally et al., 2012) or a science process skill factor (Germann, 1989). Additionally, we also find two-factor to five-factor models (Feyzioglu et al., 2012; Grube, 2010; Hammann et al., 2008a; Nowak et al., 2013), as well as models with as much as eight or 11 factors (Chang et al., 2011; Feyzioglu et al., 2012). The names of these factors are usually the same as the descriptions of the subskills and according subscales that are included in the tests. Thus, it is not surprising that we find many factor names that can be related to the four most commonly tested SR skills, *hypothesis generation*, *evidence generation*, *evidence evaluation*, and *drawing conclusions*.

Regarding the analyses that were used, it is noteworthy that most one-factor solutions resulted from exploratory principal component analyses while the models with multiple factors resulted from confirmatory factor analyses and tests of multidimensional Rasch models. Two of the models with multiple factors were tested against unidimensional models (Grube, 2010; Nowak et al., 2013). Drawing conclusions from results of factor analyses is complicated further by the fact that the interpretation of the analysis can be influenced by the assumptions of the authors. There is at least one case in which the results of a factor analysis that would allow the conclusion of a multidimensional structure are interpreted as fitting an assumed unidimensional model (Lawson, 1978). Overall, there seems to be a slightly stronger case for a multidimensional conceptualization of scientific reasoning compared to a unidimensional one: While the overall number of studies in favour of a unidimensional versus a multidimensional structure is roughly equal, the multidimensional structure is backed by more advanced statistical tools in general and by two direct model comparisons in particular. Combining this with the observation of at least one case in which the interpretation of a unidimensional model was doubtful makes multidimensionality a slightly favourable conceptualization as of now. However, a truly decisive empirical answer to the question of dimensionality cannot be given at this moment,

especially regarding the exact number of scientific reasoning factors of the slightly more probable multidimensional model.

Test context and assumptions about domain generality versus 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, 10 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.

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 seven tests using other scientific reasoning tests to establish concurrent validity and only six 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 programme, 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.

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., 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 analysed 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 analyse biology lab reports of students, helping to evaluate (amongst others) if the hypotheses are stated clearly, data are analysed 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).

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 2 decades according to the 38 scientific reasoning tests we reviewed in this article. There seems to be a shift away from considering scientific reasoning as having one single underlying cognitive ability developing in childhood and youth 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 tests might be reflective of a trend towards model-based inquiry in which questions are not just handed to students (Windschitl, Thompson, & Braaten, 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, for example, *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 has not increased in recent years, the test formats have become more diverse. Multiple-choice tests are not as common as they used to be. Instead, new test formats using virtual experiments have begun to appear. This might be reflective of the shift in science education towards practising 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. The few factor analyses that were conducted indicate that there are at least some aspects of the scientific reasoning construct that cannot be fit into a unidimensional model. However, there is also no clear alternative factorial structure. 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 that items asking for the experimental generation of evidence only apply to domains that work empirically.

The overall state of psychometric quality checks is unsatisfactory, and it is important to improve this in the future. It should become a standard procedure to check test reliability and dimensionality. In addition to the factorial structure, several other aspects of validity need more attention, too. Authors should 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, that is, 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. Furthermore, it would be informative to see if the result by Schwichow et al. (2016) showing that the effect sizes for interventions targeting the control-of-variables strategy are moderated by the test format can be replicated in other areas of scientific reasoning as well.

These shortcomings might be the reason that so far it has not been 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 (Ross, 1988), 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.

While the diversity of target groups and contexts for which the presented tests can be used does not allow us to single out one test in specific as the best scientific reasoning test,

there are certainly several heuristics that could be considered while selecting a test. Thus, 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. Developing a new test is not necessary in most instances. Instead, practitioners and researchers should start with using the list of scientific reasoning tests in Table 2. The basis of a search for a test should always be a clear idea about the conceptualization of scientific reasoning that is supposed to be tested. Next, the list of potential tests should be narrowed down. 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 straightforward 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).

Next, inspect these candidate tests in order to select the one with the best fit to the intended purpose. Of course the results from the checks of psychometric properties should play a role in the decision. However, 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. For instance, if someone wants to measure scientific reasoning in a way that is in accordance with current, broader conceptualizations of scientific reasoning (e.g., Fischer et al., 2014; NRC, 2012), it is important to look at the skills that are covered by the test. A test that only focuses on one or two skills, for example, evidence generation or drawing conclusions, would be a poor fit to these broad conceptualizations. Instead, tasks from large-scale assessment (e.g., Donovan, Lennon, et al., 2008b; Mullis et al., 2009; OECD, 2006) as well as other tests that cover a large range of skills (e.g., Nowak et al., 2013; Timmerman et al., 2011) should be considered. This is especially important taking into account that factor analyses of the presented tests did not always result in unidimensional solutions. As long as we cannot exclude the possibility that scientific reasoning is multidimensional, we should measure several aspects of the construct if we want a broad assessment of scientific reasoning because we cannot assume that a single sum score will accurately predict the performance level of subskills. However, this does not mean that tests with a narrow focus are not useful. As was pointed out in the last paragraph, if you are only interested in one aspect of scientific reasoning like the ability of students to construct an experiment, you should use a test that concentrates on this skill (Sirum & Humburg, 2011; Weld et al., 2011). Just be aware of the possibility that the test score is not only a measure of the one aspect of scientific reasoning you were interested in but that it might also be confounded with a general scientific reasoning factor (Gustafsson & Åberg-Bengtsson, 2010).

Similarly, if you only want to measure scientific reasoning in relation to one subject area, it is advisable to choose a test that is set in this specific context and aims to measure scientific reasoning in a domain-specific way (e.g., Cloonan & Hutchinson, 2011; Hammann et al., 2008b). Finally, if the theoretical assumptions that test authors make are unsound, the test should be avoided altogether. When a test uses an older theory (e.g., Inhelder & Piaget, 1958), practitioners and researchers who want to use the test should inform

themselves about criticisms of the theory and how this might influence the interpretation of the test result (Croker, 2012; Gopnik, 1996).

Especially when a test is needed for a target population that is underserved by current tests, it is possible that no test has a good fit to the intended purpose. This is one of the situations in which it is worthwhile to consider the construction of a new test. Two things should be weighed against each other in such a situation: on the one hand, the distortion of the results by the misfit of current tests and, on the other hand, the aforementioned danger of creating a narrowly fitting test that is only responsive to a very specific instance of measuring scientific reasoning but does not yield generalizable results. If a new test is created, it should be based on a clear theoretical foundation and its assumptions and psychometric properties need to be checked before its final use. Furthermore, it should be pointed out what the added benefit of the new test is compared to established tests. Even when a new test is created, it is still a good idea to also administer an established test. The comparison of the old and the new test can provide insights into the question of whether the new test is just responsive to one specific instance of scientific reasoning or if there is at least some connection with current assessments. If the results of the two tests differ widely, the reason for this difference should be explored.

Going forward, the described pragmatic approach of selecting a test 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. The importance of checking the conceptualization a test is based on highlights two important implications of this review: First, practitioners and researchers looking for a test should be weary of those tests that were not based on either a theory from the literature or an educational guideline. For these tests, it is not possible to judge whether the items really measure the construct they intend to assess. Second, the conceptualization of scientific reasoning needs further refinement. A more clearly defined conceptualization will make it easier to distinguish between tests of varying quality. Hallmarks of good conceptualizations are explanations for how skills develop, the recognition of underlying cognitive processes, and evidence to support the assumptions of the conceptualization. Studies that could support the refinement of scientific reasoning conceptualizations in this direction should have a very detailed look at scientific reasoning tasks and all the involved difficulties. Adams and Wieman (2015) conducted such a study in the area of complex problem-solving that could serve as an example for the assessment of scientific reasoning. They analysed 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.

Additionally, while there have been some advances 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 great deal 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.

Note

1. 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.

Acknowledgements

We thank Richard Shavelson and Jonathan Osborne for feedback on earlier versions of this manuscript. We are grateful to András Csanádi for his support with the coding of the tests.

Disclosure statement

No potential conflict of interest was reported by the authors.

Funding

This work was supported by the Elite Network of Bavaria [K-GS-2012-209].

Notes on contributors

Ansgar Opitz is a postdoc at the Munich Center of the Learning Sciences of the Ludwig-Maximilians-Universität München (LMU). His main research interests are the assessment of both scientific reasoning and diagnostic competences and the exploration of new statistical methods for test evaluation.

Moritz Heene is a professor of psychology in the learning sciences. His main research interests concern research methodologies, and his research deals with the question to which extent certain scientific claims can (or cannot) be derived from statistical methods commonly applied in psychology.

Frank Fischer is a professor of educational science and educational psychology at the LMU and director of the Munich Center of the Learning Sciences. His main research interests are learning with digital media, scientific reasoning and argumentation, collaborative learning, use-inspired basic research, and evidence-based practice in education.

ORCID

Ansgar Opitz http://orcid.org/0000-0002-4753-2157

References

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*, 397–419. doi: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*, 459–467. doi:10.1119/1.4913923

Azarpira, N., Amini, M., Kojuri, J., Pasalar, P., Soleimani, M., Khani, S. H., … Lankarini, 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, March-April). *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.

Baird, W. E., & Borich, G. D. (1987). Validity considerations for research on integrated-science process skills and formal reasoning ability. *Science Education*, *71*, 259–269. doi:10.1002/sce.3730710212

Baird, W. E., Shaw, E. L., Jr., & McLarty, P. (1996). Predicting success in selected events of the Science Olympiad. *School Science and Mathematics*, *96*, 85–93. doi:10.1111/j.1949-8594.1996.tb15815.x

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

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

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*, 142–174. doi:10.1080/10627197.2010.530562

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*, 169–177. doi:10.1002/tea.3660220208

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*, 1213–1233. doi:10.1007/s10763-010-9256-x

Clark, D. B., & Sampson, V. (2008). Assessing dialogic argumentation in online environments to relate structure, grounds, and conceptual quality. *Journal of Research in Science Teaching*, *45*, 293–321. doi:10.1002/tea.20216

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

Croker, S. (2012). *The development of cognition*. Boston, MA: Cengage Learning.

Dillashaw, F. G., & Okey, J. R. (1980). Test of the integrated science process skills for secondary science students. *Science Education*, *64*, 601–608. doi:10.1002/sce.3730640506

Donovan, J., Hutton, P., Lennon, M., O'Connor, G., & Morrissey, N. (2008a). *National Assessment Program – Science Literacy Year 6 school release materials, 2006*. Carlton, South Victoria, Australia: Ministerial Council on Education, Employment, Training and Youth Affairs.

Donovan, J., Lennon, M., O'Connor, G., & Morrissey, N. (2008b). *National Assessment Program – Science Literacy Year 6 report, 2006*. Carlton, South Victoria, Australia: Ministerial Council on Education, Employment, Training and Youth Affairs.

Feldon, D. F., Maher, M. A., Hurst, M., & Timmerman, B. (2015). Faculty mentors', graduate students', and performance-based assessments of students' research skill development. *American Educational Research Journal*, *52*, 334–370. doi:10.3102/0002831214549449

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*, 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. doi:10.14786/flr.v2i2.96

Fraser, B. J. (1979). *Test of Enquiry Skills [and] handbook*. Hawthorn, Victoria, Australia: 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*, 7–16. doi:10.1002/tea.3660170103

Frederiksen, N., & Ward, W. C. (1978). Measures for the study of creativity in scientific problem-solving. *Applied Psychological Measurement*, *2*, 1–24. doi:10.1177/014662167800200101

Germann, P. J. (1989). The Processes of Biological Investigations Test. *Journal of Research in Science Teaching*, *26*, 609–625. doi:10.1002/tea.3660260706

Gilmore, J., Vieyra, M., Timmerman, B., Feldon, D., & Maher, M. (2015). The relationship between undergraduate research participation and subsequent research performance of early career STEM graduate students. *The Journal of Higher Education*, *86*, 834–863. doi:10.1353/jhe.2015.0031

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*, 521–563. doi:10.1080/10508406.2013.837391

Gopnik, A. (1996). The post-Piaget era. *Psychological Science*, *7*, 221–225. doi:10.1111/j.1467-9280.1996.tb00363.x

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*, 364–377. doi:10.1187/cbe.12-03-0026

Grube, C. (2010). *Kompetenzen naturwissenschaftlicher Erkenntnisgewinnung: Untersuchung der Struktur und Entwicklung des wissenschaftlichen Denkens bei Schülerinnen und Schülern der Sekundarstufe I* [Epistemic competencies in science: Investigation of structure and development of scientific thinking for secondary school pupils] (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.

Hammann, M., Phan, T. H., & Bayrhuber, H. (2008a). Experimentieren als Problemlösen: Lässt sich das SDDS-Modell nutzen, um unterschiedliche Dimensionen beim Experimentieren zu messen? [Experimentation as problem-solving: Can the SDDS model be used to measure different dimensions in experimentation?]. In M. Prenzel, I. Gogolin, & H.-H. Krüger (Eds.), *Kompetenzdiagnostik* [Competence diagnosis] (pp. 33–49). Wiesbaden, Germany: VS Verlag für Sozialwissenschaften.

Hammann, M., Phan, T. T. H., Ehmer, M., & Grimm, T. (2008b). Assessing pupils' skills in experimentation. *Journal of Biological Education*, *42*, 66–72. doi:10.1080/00219266.2008.9656113

Heene, M. (2007). *Konstruktion und Evaluation eines Studierendenauswahlverfahrens für Psychologie an der Universität Heidelberg* [Construction and evaluation of a student admission test for psychology at the university of Heidelberg] (Doctoral dissertation). Retrieved from http://archiv.ub.uni-heidelberg.de/volltextserver/7727/1/Diss_Text7_final_published.pdf

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

Klahr, D., & Dunbar, K. (1988). Dual space search during scientific reasoning. *Cognitive Science*, *12*, 1–48. doi:10.1207/s15516709cog1201_1

Klos, S. (2009). *Kompetenzförderung im naturwissenschaftlichen Anfangsunterricht – Der Einfluss eines integrierten Unterrichtskonzepts* [Training competencies in science education for beginners – The influence of an integrated curriculum]. Berlin, Germany: Logos Verlag.

Klos, S., Henke, C., Kieren, C., Walpuski, M., & Sumfleth, E. (2008). Naturwissenschaftliches Experimentieren und chemisches Fachwissen – Zwei verschiedene Kompetenzen [Experimenting in science and chemical content knowledge – Two separate competencies]. *Zeitschrift für Pädagogik*, *54*, 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*, 327–336. doi:10.1111/cdev.12298

Lawson, A. E. (1978). The development and validation of a classroom test of formal reasoning. *Journal of Research in Science Teaching*, *15*, 11–24. doi:10.1002/tea.3660150103

Lawson, A. E., Alkhoury, S., Benford, R., Clark, B. R., & Falconer, K. A. (2000a). What kinds of scientific concepts exist? Concept construction and intellectual development in college biology. *Journal of Research in Science Teaching*, *37*, 996–1018. doi:10.1002/1098-2736(200011)37:9<996::AID-TEA8>3.0.CO;2-J

Lawson, A. E., Clark, B., Cramer-Meldrum, E., Falconer, K. A., Sequist, J. M., & Kwon, Y.-J. (2000b). Development of scientific reasoning in college biology: Do two levels of general hypothesis-

testing skills exist? *Journal of Research in Science Teaching*, *37*, 81–101. doi:10.1002/(SICI)1098-2736(200001)37:1<81::AID-TEA6>3.0.CO;2-I

Lazonder, A. W., & Harmsen, R. (2016). Meta-analysis of inquiry-based learning: Effects of guidance. *Review of Educational Research*, *86*, 681–718. doi: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*, 497–521. doi:10.1002/tea.10034

Linn, M. C., Pulos, S., & Gans, A. (1981). Correlates of formal reasoning: Content and problem effects. *Journal of Research in Science Teaching*, *18*, 435–447. doi:10.1002/tea.3660180507

Linn, M. C., & Rice, M. (1979). A measure of scientific reasoning: The Springs task. *Journal of Educational Measurement*, *16*, 55–58. doi:10.1111/j.1745-3984.1979.tb00087.x

Linn, M. C., & Swiney, J. F. (1981). Individual differences in formal thought: Role of expectations and aptitudes. *Journal of Educational Psychology*, *73*, 274–286. doi:10.1037/0022-0663.73.2.274

Livermore, A. H. (1964). The process approach of the AAAS commission on science education. *Journal of Research in Science Teaching*, *2*, 271–282. doi:10.1002/tea.3660020403

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, Lynch School of Education, 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, Lynch School of Education, Boston College.

Mayer, D. (2012). *Die Modellierung des wissenschaftlichen Denkens im Grundschulalter* [Modelling of scientific thinking in elementary school age] (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. doi:10.1016/j.learninstruc.2013.07.005

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. doi:10.1016/j.sbspro.2012.09.434

Molitor, L. L., & George, K. D. (1976). Development of a test of science process skills. *Journal of Research in Science Teaching*, *13*, 405–412. doi:10.1002/tea.3660130504

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, Lynch School of Education, Boston College.

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

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

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

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, The Netherlands: Springer.

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*, 182–188. doi:10.1080/00219266.2013.822747

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

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

Organisation for Economic Co-operation and Development. (2009). *PISA 2006 technical report*. Paris, France: Author.

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

Padilla, M. J., Okey, J. R., & Dillashaw, F. G. (1983). The relationship between science process skill and formal thinking abilities. *Journal of Research in Science Teaching*, *20*, 239–246. doi:10.1002/tea.3660200308

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* [IQB state comparison 2012: Mathematical and science competencies at the end of stage 1 of secondary education]. Münster, Germany: 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. doi:10.1016/j.edurev.2015.02.003

Perkins, D. N., & Salomon, G. (1989). Are cognitive skills context-bound? *Educational Researcher*, *18*(1), 16–25. doi:10.3102/0013189X018001016

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*, 5–21. doi:10.1080/0263514042000187511

Roberts, R., & Gott, R. (2006). Assessment of performance in practical science and pupil attributes. *Assessment in Education: Principles, Policy & Practice*, *13*, 45–67. doi:10.1080/09695940600563652

Ross, J. A. (1988). Controlling variables: A meta-analysis of training studies. *Review of Educational Research*, *58*, 405–437. doi:10.3102/00346543058004405

Ross, J. A., & Maynes, F. J. (1983). Development of a test of experimental problem-solving skills. *Journal of Research in Science Teaching*, *20*, 63–75. doi:10.1002/tea.3660200107

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. doi:10.1002/tea.21303

Schwichow, M., Croker, S., Zimmerman, C., Höffler, T., & Härtig, H. (2016). Teaching the control-of-variables strategy: A meta-analysis. *Developmental Review*, *39*, 37–63. doi:10.1016/j.dr.2015.12.001

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. doi:10.1080/00091380309604739

Shaw, T. J. (1983). The effect of a process-oriented science curriculum upon problem-solving ability. *Science Education*, *67*, 615–623. doi:10.1002/sce.3730670510

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

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

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.

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

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

Timmerman, B. C., Feldon, D., Maher, M., Strickland, D., & Gilmore, J. (2013). Performance-based assessment of graduate student research skills: Timing, trajectory, and potential thresholds. *Studies in Higher Education*, *38*, 693–710. doi:10.1080/03075079.2011.590971

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*, 509–547. doi: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*, 413–423. doi:10.1177/001316448104100220

Tobin, K. G., & Capie, W. (1982). Relationships between formal reasoning ability, locus of control, academic engagement and integrated process skill achievement. *Journal of Research in Science Teaching*, *19*, 113–121. doi:10.1002/tea.3660190203

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.

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*, 11–29. doi:10.1111/j.1745-3984.1980.tb00811.x

Weld, J., Stier, M., & McNew-Birren (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. Retrieved from http://www.jstor.org/stable/42992885

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. Retrieved from http://www.jstor.org/stable/42993885

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*, 254–284. doi:10.3102/0091732X09351544

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*, 941–967. doi:10.1002/sce.20259

Wu, M., Donovan, J., Hutton, P., & Lennon, M. (2008). *National Assessment Program – Science Literacy Year 6 technical report, 2006*. Carlton, South Victoria, Australia: Ministerial Council on Education, Employment, Training and Youth Affairs.