Journal for Educational Research Online
Journal für Bildungsforschung Online
Volume 5 (2013), No. 2, 110–138
© 2013 Waxmann

Inga Hahn, Katrin Schöps, Silke Rönnebeck, Maike Martensen, Sabine Hansen, Steffani Saß, Inger Marie Dalehefte & Manfred Prenzel

Assessing scientific literacy over the lifespan – A description of the NEPS science framework and the test development

Abstract

The first part of the paper describes the science framework of the National Educational Panel Study (NEPS) that forms the basis for assessing scientific literacy over a person's lifespan. The framework and its definition of scientific literacy are influenced by the Programme for International Student Assessment (PISA), by the American Association for the Advancement of Science (AAAS) and by the German educational standards for the end of Grade 10. All of these sources claim that scientific literacy is important for everyone: It forms a basis for general education, has to be applicable to everyday situations and is a source for lifelong learning. Thus, the contexts and components providing the foundation for item development and for measuring scientific literacy had to be chosen accordingly. This paper presents a selection of contexts and content areas that meet this demand and that, at the same time, allow NEPS to be theoretically and methodologically linked to other national and international large-scale assessments.

The second part of the paper is concerned with the process of item selection. Since NEPS aims to measure scientific literacy over the lifespan, tests for a variety of different age groups have to be developed. Psychometric properties of pilot study tests for children in kindergarten, Grade 6 and Grade 9 are presented. The

Dr. Inga Hahn (corresponding author) · Dr. Katrin Schöps · Dr. Silke Rönnebeck · Dr. Maike Martensen · Dr. Sabine Hansen · Dr. Steffani Saß · Dr. Inger Marie Dalehefte, Leibniz Institute for Science and Mathematics Education (IPN) at the University of Kiel, Olshausenstraße 62, 24118 Kiel, Germany
e-mail: hahn@ipn.uni-kiel.de
schoeps@ipn.uni-kiel.de
roennebeck@ipn.uni-kiel.de
martensen@ipn.uni-kiel.de
shansen@ipn.uni-kiel.de
sass@ipn.uni-kiel.de
dalehefte@ipn.uni-kiel.de

Prof. Dr. Manfred Prenzel, TU Munich School of Education, Arcisstraße 21, 80333 Munich, Germany
e-mail: manfred.prenzel@tum.de

paper concludes with an outlook of further ways to validate the preliminary findings and of linking tests for different age groups.

Keywords
Scientific literacy; Test development; Panel study; Lifespan

Die Erfassung naturwissenschaftlicher Kompetenz über die Lebensspanne – eine Beschreibung der Rahmenkonzeption und der Entwicklung des NEPS-Naturwissenschaftstests

Zusammenfassung
Der erste Teil des Artikels beschreibt die Rahmenkonzeption naturwissenschaftlicher Kompetenz des NEPS – Bildungsverläufe in Deutschland. Sie bildet die Grundlage für die Messung naturwissenschaftlicher Kompetenz über die Lebensspanne. Das Rahmenkonzept und die Definition naturwissenschaftlicher Kompetenz sind durch das Programme for International Student Assessment (PISA), die Standards der American Association for the Advancement of Science (AAAS) sowie durch die Deutschen Bildungsstandards für den mittleren Bildungsabschluss beeinflusst. Gemäß diesen Quellen ist naturwissenschaftliche Kompetenz wichtig für jedermann. Sie ist eine der Grundlagen der Allgemeinbildung und des lebenslangen Lernens. Die Kontexte und Inhaltsbereiche, die die Basis für die Aufgabenentwicklung und für die Messung naturwissenschaftlicher Kompetenz bilden, wurden diesen Vorgaben entsprechend ausgewählt. Der Artikel präsentiert die Kontexte und Inhaltsbereiche, die diese Anforderungen erfüllen und die es darüber hinaus erlauben, NEPS theoretisch und methodisch mit anderen nationalen und internationalen Large-Scale Assessments zu verbinden.

Der zweite Teil des Artikels widmet sich dem Prozess der Itementwicklung. Da NEPS auf die Messung naturwissenschaftlicher Kompetenz über die Lebensspanne abzielt, müssen Testinstrumente für eine Vielzahl verschiedener Altersgruppen entwickelt werden. Die psychometrischen Eigenschaften der Großpilottests für Kindergartenkinder, sowie für Kinder der sechsten und neunten Klasse werden berichtet. Der Artikel schließt mit einem Ausblick auf weitere geplante Studien zur Validierung der bisherigen Ergebnisse und zur Verlinkung der verschiedenen Tests über die Altersstufen.

Schlagworte
Naturwissenschaftliche Kompetenz; Testentwicklung; Panelstudie; Lebensspanne

1. Introduction

In modern societies daily life is strongly influenced and determined by science. Science helps us to understand the world and the ways in which our world is changing due to scientific and technological progress. It also empowers us to understand, evaluate and address daily problems, but also issues of social and worldwide significance. Thus, it is widely agreed that science education is important for everyone (Osborne & Dillon, 2008) and that the acquisition of basic science competencies is a central goal of schooling. The question remains, however, about how young children develop an understanding of science, how this understanding can be promoted and whether the basic scientific knowledge that is later imparted in school can provide a basis for lifelong learning.

Up to now, no longitudinal large-scale studies that could answer these questions have been carried out in Germany. The two big studies measuring science competencies in Germany – Trends in International Mathematics and Science Study (TIMSS) and Programme for International Student Assessment (PISA) – only provide cross-sectional data and only address specific age groups. Therefore, these studies are not suitable for assessing the development of skills and competencies over the lifespan and for answering corresponding research questions. This gap in educational research in Germany is filled by the National Educational Panel Study (NEPS) that aims to study the development of various aspects of education over the lifespan. The idea of the panel is to accompany test persons throughout their life and to repeatedly assess different competencies at different stages of their life. The intervals between measurements depend on the competency measured and on the test persons' stages of life. From kindergarten to the age of 18, scientific literacy is assessed every two to three years.

A main focus of NEPS is placed on the development of competencies in mathematics, reading (including listening), science, and information and communications technology (ICT). In 2009, the NEPS science group started its work by developing a theoretical framework defining the structure and content of scientific literacy for age groups ranging from kindergarten to retirement age.

The following paper provides a definition of what is meant in NEPS by the term *scientific literacy* and gives an overview of the NEPS assessment framework. Based on this framework, item and test development processes are described and illustrated by item examples. First results of the test development and validation are presented for different pilot studies in kindergarten, Grade 6 and Grade 9.

2. Theoretical Background

2.1 The NEPS Science Framework

The definition of scientific literacy used by NEPS includes aspects of the *concept of competence* as defined by Weinert (2001) and the concepts of *scientific literacy* developed by the American Association for the Advancement of Science (AAAS, 1993, 2009) and by PISA (Bybee, McCrae, & Laurie, 2009; Bybee & *PISA 2006 Science Expert Group*, 2009; Bybee, 1997a, 1997b; Gräber, Nentwig, Koballa, & Evans, 2002; OECD, 2006; Prenzel & Seidel, 2008; Prenzel et al., 2007; Prenzel, 2000).

According to Weinert (2001), competencies should be considered as cognitive problem solving skills that are either inherited or acquired by a person. Connected to these skills is a person's motivational, volitional and social disposition to solve problems in different situations. This disposition is addressed by other working groups within NEPS and is therefore not part of the NEPS definition of scientific literacy.

Weinert's definition of the cognitive components of competency is complemented by Klieme's and Leutner's view that competencies are context-specific achievement dispositions which functionally refer to domain-specific situations and demands (Klieme & Leutner, 2006). A person who acts competently uses his or her knowledge actively and is able to cope with the demands of everyday life situations. Hence, success is not accidental but based on a latent trait which enables a person to adequately master unknown situations (Klieme et al., 2004).

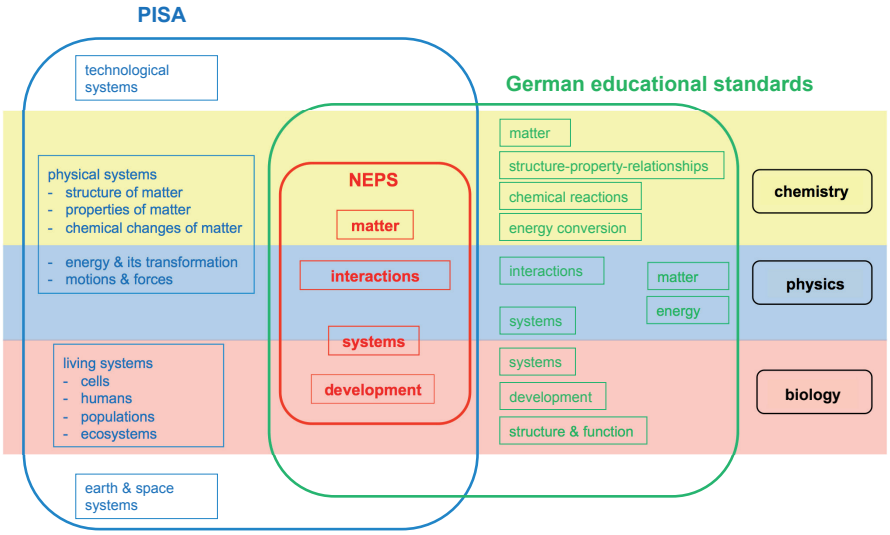
According to Laugksch (2000), the term scientific literacy is widely used but often with different definitions and conceptualization. At present, two of the most commonly used concepts of scientific literacy are the ones outlined in the PISA 2006 scientific literacy framework and in the *Benchmarks for Science Literacy* of the American Association for the Advancement of Science (AAAS, 1993, 2009; Bybee, McCrae, & Laurie, 2009; Bybee & *PISA 2006 Science Expert Group*, 2009; Bybee, 1997a, 1997b; Gräber et al., 2002; OECD, 2006; Prenzel & Seidel, 2008; Prenzel et al., 2007; Prenzel, 2000). Both concepts comply with the idea that a basic understanding of scientific concepts and processes is relevant for everyone in his or her daily life and forms the basis for lifelong learning. Rather than focusing on memorized knowledge, scientific literacy reflects the ability to apply one's existing scientific knowledge in different everyday life contexts and situations. This broad idea of literacy recognizes the importance and relevance of the competencies, knowledge, methods, and values that define the scientific disciplines and that are considered to be of great importance for an actively participating citizen (see also Klahr, 2000; Klahr & Dunbar, 1988; Mayer, 2007). Our rapidly changing and developing society increasingly demands scientific literacy in order to understand and make use of technological innovations, to adequately face environmental challenges (e.g., climate changes), and to reflect on one's own actions as a responsible citizen. Similarly, Miller (1983) postulated that scientific literacy in a contemporary situ-

ation consists of three dimensions: (a) an understanding of the norms and methods of science (i.e., the nature of science); (b) an understanding of key scientific terms and concepts (i.e., science content knowledge); and (c) an awareness and understanding of the impact of science and technology on society. However, when it comes to defining the key scientific concepts a scientifically literate person should know or master, one also finds broad disagreement (Shwartz, Ben-Zvi, & Hofstein, 2006). Hence, NEPS used a pragmatic approach for selecting the key concepts that were to be included in its framework.

In the first step, the commonalities of the PISA 2006 scientific literacy framework (OECD, 2006), the *Benchmarks for Science Literacy* of the American Association for the Advancement of Science (AAAS, 1993, 2009) and the German educational standards for the end of Grade 10 (*Bildungsstandards für den Mittleren Schulabschluss;* KMK, 2005a, 2005b, 2005c) were identified. As the latter serve as educational standards for everybody who successfully passes through the German education system and gains a secondary school certificate, they have to be included in every national framework that aims to assess scientific literacy. Additionally, the choice of PISA and the German educational standards as reference points for the NEPS science framework complies with the requirement stated by the German Ministers of Education and Cultural Affairs of the Federal States that national and international large-scale assessments should be theoretically and methodologically linked.

In the second step, the commonalities of the PISA 2006 scientific literacy framework, the *Benchmarks for Science Literacy* of the American Association for the Advancement of Science and the German educational standards were reduced to contexts and contents that show a substantial overlap and that could be assessed in the rather limited 30 minutes of testing time that is available for the NEPS science tests. Figure 1 gives an overview of the resulting content overlap between PISA, the German educational standards and NEPS in the *knowledge of science* (KOS) area. In the center of Figure 1, the NEPS components which measure knowledge of science are presented. The horizontal yellow, blue and pink sections show the concepts covered by PISA and the German educational standards which are incorporated into the corresponding NEPS components. Where two NEPS components are given (e.g., *interactions* and *systems* in the blue section), the concepts of PISA and the German educational standards can be partly aligned to both of the NEPS components.

Figure 1: Overview of the KOS content overlap between PISA, the German educational standards and NEPS

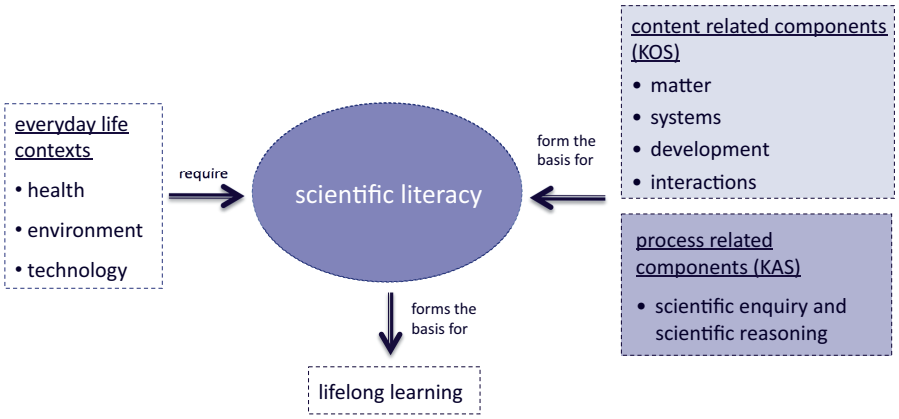


Similar to the definition used by PISA (OECD, 2006), the NEPS framework distinguishes between KOS or rather knowledge of basic scientific concepts and facts, and *knowledge about science* (KAS) or the understanding of scientific processes (see also Hodson, 1992). KOS is divided into the content-related components *matter*, *systems*, *development* and *interactions*. KAS is divided into the process-related components *scientific enquiry* and *scientific reasoning* (called *scientific explanations* in PISA). Hypothesizing, planning experiments and analyzing data are concepts of the German educational standards which have been incorporated into the NEPS component *scientific enquiry*. The concepts of drawing conclusions, of interpreting results and of identifying and dealing with measurement errors are part of the NEPS component *scientific reasoning*.

KOS and KAS are implemented in three selected everyday life contexts: *health, environment* and *technology*. Since the German educational standards do not explicitly mention contexts, the choice of the NEPS contexts was based on the PISA framework and on the idea of choosing areas that are generally agreed to be of lifelong significance. Since this was tested by experts, students and PISA managers around the world during the PISA test developments of 2000, 2003 and 2006 (Fensham, 2009), and since NEPS is required to show a connection to other national and international large-scale assessments, the NEPS expert group decided to rely on the PISA choice of contexts.

The NEPS contexts do not explicitly consider the PISA perspective of personal, social and global situations. These situations occur implicitly but cannot be analyzed separately due to the limited number of items representing each situation. Therefore, they are not important for the NEPS framework that is given in Figure 2.

Figure 2: Contexts and components of the NEPS framework for measuring scientific literacy



2.1.1 The content related components (knowledge of science)

There is a vast amount of scientific aspects that justifiably could have been included in the NEPS framework. Therefore, it was necessary to prioritize and structure the content related components for the assessment of knowledge of science as described in section 2.1. Finally, the content related components *matter*, *systems*, *development* and *interactions* were selected for NEPS as it is widely agreed that they cover key aspects of science and are also central to the reference frameworks mentioned in section 2.1 (Bybee, McCrae, & Laurie, 2009; Bybee & *PISA 2006 Science Expert Group*, 2009; Bybee, 1997a, 1997b; Council of Ministers of Education, 1997; Gräber et al., 2002; National Research Council, 1998; OECD, 2006; Prenzel & Seidel, 2008; Prenzel et al., 2007; Prenzel, 2000; KMK, 2005a, 2005b, 2005c; AAAS, 1993, 2009).

2.1.1.1 Matter. Matter is the foundation of all natural processes and hence is relevant for all natural sciences. The ideas of the discontinuity as well as the conservation of matter are two of the key concepts in science according to the AAAS *Benchmarks for Science Literacy* (AAAS, 1993, 2009) and are also taken up in the German educational standards for chemistry and physics (KMK, 2005b, 2005c) at the end of Grade 10. The scientific understanding of atoms and molecules requires combining two closely related ideas. All substances are composed of invisible particles that belong to a limited number of elements. Combining the particles differently results in millions of materials with different properties. An understanding of atomic and molecular theory has very important implications for an understanding of science and of how the world works. It provides the basis for understanding macroscopic phenomena such as melting glaciers, global warming and the development of pandemics. Since the understanding and conceptualization of a theory

of matter changes throughout the years students spend at school and beyond this time, the concept of matter is a concept which is important to measure over a lifespan (Carey, 1991; Smith, Wiser, Anderson, & Krajcik, 2006).

2.1.1.2 Systems. Systems belong to those "important themes [that] pervade science, mathematics, and technology and appear over and over again [...]. They are ideas that transcend disciplinary boundaries and prove fruitful in explanation, in theory, in observation, and in design" (AAAS, 2007, p. 90). A sound understanding of systems requires the ability to distinguish and switch between different levels of a system in order to assess it from different perspectives. The interactions between different components need to be understood in order to gauge and predict the effects certain changes will have on a system (e.g., Ahlgren & Rutherford, 1991; Hurd, 1998; Millar & Osborne, 1998; UNESCO, 1993). The NEPS framework distinguishes between biological and technological systems:

a) Biological systems
An understanding of biological systems requires, for example, knowledge of the different kinds of relationships that exist between and within organisms and components of a system, knowledge of the *variety* of physical conditions that organisms must cope with, and knowledge of the kind of environments created by the interaction of organisms with one another and their physical surroundings, hence, knowledge of the complexity of such systems. The AAAS *Benchmarks for Science Literacy* (AAAS, 2009) and the German educational standards for biology (KMK, 2005a) emphasize that an understanding of different systems is essential in order to understand the world around us. Biological systems span from the micro-level, with cells where energy is transformed and information is processed and passed on, to the macro-level, with the global systems of our biosphere like biogeochemical cycles and their energy flow. Topics such as biotoxins, epidemics, global warming and sea level rise can only be understood with a comprehensive understanding of the affected systems.

b) Technological systems
In contrast to natural systems, technological systems are mostly constructed from the laws of system theory – whereas forests or the carbon cycle, for instance, can be interpreted systemically, technological systems are initially constructed from these principles. They play a major role in our daily life with respect to supply and maintenance cycles like water or electricity supply and waste disposal. An important aspect is the ability to control and influence the system in order to adapt the supply to people's needs. Understanding the components, interactions and principles of technological systems requires a basic systemic knowledge of concepts like control elements, set-actual comparisons, disturbance variables, feedback, energy conservation, constraints and the need for making trade-offs. According to

the European Commission, a basic understanding of technology belongs to the eight key competences for lifelong learning (European Commission, 2007). All of the key competences emphasize critical thinking, creativity, initiative, problem solving, risk assessment, and decision making and all of these aspects play a major role in technology education.

2.1.1.3 Development. Another key concept in biology is development (AAAS, 1993, 2009; KMK, 2005a; Millar & Osborne, 1998; Trefil, 2008). It includes not only the development of organisms and different species but also heredity and consequently the biodiversity of life on earth. Even young children want to know where they come from and why people who are related with each other show certain similarities. For an understanding of the processes involved in development and heredity, knowledge of different biological sub-disciplines such as cell biology, genetics and physiology is necessary (KMK, 2005a). An understanding of cells and molecules is necessary in order to understand what genetic information is, how it is passed on from one generation to the next and, subsequently, how species evolve.

2.1.1.4 Interactions. Interactions are part of the AAAS *Benchmarks for Science Literacy* (AAAS, 1993, 2009) within the concept of motions. In our daily life, we are surrounded by interactions. It is one of nature's basic underlying principles that two bodies that get in contact with each other interact – if a chair is pushed, it moves. The NEPS framework focuses on two types of interactions, mechanical interactions (leading to deformation or changes in the state of motion) and interactions between waves (i.e., sound waves or electromagnetic radiation) and matter. An understanding of interactions requires, for example, knowledge of forces and motions, of the properties of different kinds of waves and of the ways in which waves can interact with matter. Interactions were chosen as one of the components assessed in NEPS because we not only encounter them regularly but also because they have a high relevance in our daily lives (examples range from sun protection and medical x-ray diagnostics to hazards like earthquakes and nuclear accidents). Moreover, as one of the key concepts in the German educational standards for physics (KMK, 2005c), they are regarded as a framework for structuring learning during schooling and for combining new experiences with existing knowledge. They thus allow students to develop knowledge that is compatible and forms the basis for lifelong learning in a rapidly changing world (Senatsverwaltung für Bildung, Jugend und Sport Berlin, 2006).

2.1.2 The process related components (knowledge about science)

Scientific enquiry and scientific reasoning. The process related components of the NEPS science framework focus on understanding (the nature of) science as a curiosity-based human activity or endeavor with great potential but also with lim-

itations (KMK, 2005a, 2005b, 2005c; Bybee, McCrae, & Laurie, 2009; Bybee & *PISA 2006 Science Expert Group*, 2009; Bybee, 1997a, 1997b; Gräber et al., 2002; OECD, 2006; Prenzel & Seidel, 2008; Prenzel et al., 2007; Prenzel, 2000; AAAS, 1993, 2009). If a person wants, for example, to assess the scientific content and merit of information spread by the media, he or she needs knowledge of science (as described in section 2.1.1). However, one also needs knowledge about science as a process in order to gauge, for example, whether a question can in fact be answered by scientific enquiry or how trustworthy the presented results of a study are. The fact that most individuals acquire knowledge not through their own scientific investigations but through resources such as the media, libraries and the Internet puts a particular emphasis on the last point. The process related component of *scientific enquiry* relates to the understanding of science as a process consisting, for example, of posing scientific questions, hypothesizing, planning experiments (including strategies for controlling variables), analyzing data and drawing conclusions while referring to the hypotheses. These components are widely understood to be the central components of scientific processes (Koslowski, 1996; Klahr, 2000; Mayer, 2007).

A scientifically competent person should be able to draw evidence-based conclusions after selecting and evaluating information and data, and should be able to notice when the information is not sufficient or when measurement errors threaten the reliability and validity of a study. Scientific studies generate data that are often acquired by using measurement tools. As these data are the foundation for testing hypotheses and eventually for the results of a study, one has to understand how a study is designed, how the data have been collected and analyzed, what measurement tools have been used and which standards the measurements comply with (Masnick & Klahr, 2003). This knowledge is important in order to identify and deal with measurement errors (Chinn & Brewer, 1998; Zimmerman, 2007). The concepts of interpreting results and of identifying and dealing with measurement errors are part of the NEPS component *scientific reasoning*.

2.1.3 The three NEPS contexts

2.1.3.1 Health. The health context is of particular individual relevance since our health strongly influences our quality of life. Almost every day we hear about health related topics such as diseases, epidemics, nutrition and accidents in the media and face decisions concerning our own health.

2.1.3.2 Environment. In an industrialized country like Germany, every human uses large amounts of natural resources and produces tons of waste during his or her life. Therefore, environmental issues are of major importance for each individual but also for our society in order to maintain a world in which human beings can live. The disposal of waste, global change, overpopulation and pollution are only some of the major challenges our society faces in the 21st century.

2.1.3.3 Technology. Our modern world is highly technologized and – from a technological perspective – is also rapidly changing. New technological developments enter the market and people's lives every day. Consequently, most people cannot imagine their daily routines without technological devices such as cars, household appliances, mobile phones or satellite TV.

For the NEPS tests, these three contexts provide the frame in which the test items are embedded.

2.2 The NEPS Science Test

2.2.1 Test and item structure

Similar to the PISA approach, the NEPS test items are organized in units (testlets). The units are developed based on a combination of the contexts and components described in the science framework. Each item is developed in a way that it measures one component (KAS or KOS) and has one of the named contexts as a background. Due to the unit structure and the limited number of items that can be administered in half an hour of testing time, not all context-component combinations could be tested. Thus, for the purpose of validation, every component is measured in at least two contexts. The allocation of components to contexts was first of all content-related and second of all based on the idea that each context should be represented equally with a nearly equal number of items. Since only one science score per person is reported due to the small number of items in each cell in Table 1, it is important that the structure of contexts and components is the same or nearly the same for each age cohort in order to obtain comparable results.

A unit starts with a stimulus consisting of a text that can be supplemented by tables, graphs or images. This unit stimulus "tells a story" to set the stage and provide the information for the test items. This allows the items to explore a topic from different perspectives and to assess multiple aspects of performance. The stimulus is typically followed by two to four test items – sometimes further information is provided in the item stem. Each unit assesses scientific literacy within either one of the four KOS- or one of the two KAS-components, respectively, and is situated in one of the three NEPS contexts. Hence, each unit covers one context and one component. For each age cohort, the final science test in the main study consists of 23–26 items corresponding to 30 minutes of testing time. As a consequence of the large number of tests each person has to take during their participation in NEPS, the testing time per competency domain had to be minimized in order to avoid or at least to reduce panel mortality. Thus, the relatively short testing time represents a compromise between the time needed to assess a competency and the testing time participants can be expected to accept and endure.

The test items for all tests are either constructed as simple or as complex multiple choice items. The complex multiple choice items are constructed in the special

form of multiple true-false (MTF) items. This format avoids the disadvantages of the complex multiple choice format such as, for example, items with lower discrimination (and thereby lower reliability) being produced and construct irrelevant difficulty being imposed on the items (Haladyna, 1992; Haladyna & Downing, 1989), and is very effective in terms of reliability and validity (Frisbie, 1992; Albanese & Sabers, 1988). Table 1 gives an overview of the number of items used for the different context-component combinations for the final tests.

Table 1: Numbers of items for the different context-component combinations intended for the final test

		Context			
Component		Environment	Health	Technology	Total
Development	KOS	3–4			3–4
Interactions	KOS		3–4		3–4
Matter	KOS	5–6			5–6
Systems	KOS	3		3	6
Scientific enquiry	KAS	3			3
Scientific reasoning	KAS	3			3
				Total item number	23–26

Note. The number of items differs slightly because of the unit structure of the test, the item selection of the pretest and due to special test requirements for the younger age groups (children in kindergarten and first grade).

The item structure generally is the same for all age groups (see Figure A2). However, the item format for children in kindergarten (and Grade 1) slightly differs from the format for all other age groups because of the pre-schoolers' limited abilities to read and their difficulties in staying focused for longer periods of time. In contrast to the text-based answer categories for the older age cohorts, the children in kindergarten have to choose from pictures which represent the multiple choice answer categories. In addition to the special item structure, the science test in kindergarten is also characterized by a special setting, with an interviewer testing only one child at a time, and by the fact that the items are embedded in a story. The interviewer tells the children a story in which the children accompany the two main characters Paul and Lena through a summer party in kindergarten. Throughout the story, Paul and Lena come across situations in which different scientific "problems" have to be solved. The children are asked to help solve these problems and to answer the interviewer's questions by choosing an answer from certain pictures (see Figure A1).

2.2.2 Test development procedure

The test development procedure consists of different steps. In a first step approximately 60–75 items are developed. For the younger age cohorts (from kindergarten to Grade 6) these items are pre-tested in cognitive lab studies. Due to the lack of science tests for these age groups, cognitive lab studies are essential for eliminating comprehension problems and ensuring that the test material and the item difficulties are appropriate for the respective age group. After revising the items, they are pre-piloted at local schools and kindergartens. A minimum number of 100 test persons per item are tested to allow for an item response theory (IRT) analysis of the data (Linacre, 1994). Parallel to the pre-piloting, an expert group consisting of psychologists, educational researchers and natural scientists is asked to review the items with respect to their content-related correctness, their age-appropriateness, their proximity to national and international educational standards and their fit to the NEPS science framework. After analyzing the data and taking into account the feedback of the experts, approximately 40–50 items are selected for the pilot study. If necessary, the selected items are revised according to the pre-pilot results and the experts' comments. The data from the pilot study is analyzed using IRT models. Based on the results of these analyses, 23–26 items are selected for the final test. All studies that take place before the main study serve the purpose of testing and selecting appropriate items for a final test. During the main study, this final test is administered to assess the target persons' scientific literacy. Thus, the pilot and main studies not only differ in their technical properties, like the number of items and testing time, but also in a theoretical way – concerning the kind of research questions they want to answer.

In order to serve the purpose of selecting the best items for the main test, the following research questions have to be answered by the pilot studies presented in this paper:

1) Do the statistical properties of the items comply with internationally accepted standards?
   a. Do the items reach a discrimination > .30?
   b. Do the item difficulties cover the range of the target persons' abilities?
   c. Do the items fulfill the Rasch criterion of item homogeneity?
   d. Is there evidence for differential item functioning?
2) Does the test measure scientific literacy in a reliable and valid way?
   a. Does the test show acceptable reliability values?
   b. Does the hypothetical competency structure apply to the data (internal validity)?
   c. Does the science test show acceptable correlations with the measures of interest in science (external validity)?

3. Method

The analyses presented in this paper focus on the empirical characteristics of individual items and the results of the pilot tests. In order to evaluate the quality of the test, different criteria are taken into account. These criteria include parameters at the item level as well as parameters that are related to test quality such as reliability and validity aspects (internal and external). Internal validity is analyzed by assessing the dimensionality of the test (one-dimensional versus two-dimensional) and external validity by computing correlations between a person's scientific literacy and their interest in science. To date, science tests for kindergarten, for Grade 6 and Grade 9 have been developed. Therefore, this paper will focus on describing the test development and reporting the corresponding results for these three age groups.

3.1 Procedure and participants

The NEPS testing procedures in Grades 6 and 9 are very similar. The process of drawing samples included whole school classes. Tests are administered in a classroom situation where all of the students who got their parents' permission are tested. In the pilot studies, the testing time is one hour, whereas in the main study it is limited to half an hour. In addition to the competency tests, all participants fill out a questionnaire with background information about, for example, their family, their social and educational background, their learning environment and their leisure activities.

For the kindergarten tests, a different test setting was used. At that age, meaningful data can only be obtained in a one-to-one interview situation because the children are not yet able to read. The test items are read to them and the answers are given using picture-based answer categories (see Figure A1). Information on the child's background is given by the child's parents and by the kindergarten teachers. In the pilot study, each participating child in the kindergarten cohort is tested on two consecutive days. Each test lasts 30 minutes which amounts to 60 minutes of testing time. In the main study, each child is tested for 30 minutes on a single day. The testing time is thus the same as for the school cohorts (see Table 2).

Table 2: Overview of age range, test procedures, sample sizes and testing time per study

Sample	Age range	Procedure	Pilot study		Main study	
			Achieved sample size (*n*)	Testing time in minutes	Intended sample size (*n*)	Testing time in minutes
Kindergarten	4 to 5	One-to-one interview, picture-based item stem and picture based answers (multiple-choice and multiple-true-false item format)	132 (f = 67, m = 65)	60	3,000	30
Grade 6	10 to 13	Group testing (multiple-choice and multiple-true-false item format)	369 (f = 196, m = 173)	60	5,313	30
Grade 9	14 to 17	Group testing (multiple-choice and multiple-true-false item format)	182 (f = 93, m = 89)	60	12,500	30

Note. f = female; m = male.

3.2 Analyses and critical parameters

The analysis of competency data in NEPS is based on IRT. It makes it possible to refer a test person's manifest behavior (the item response) to the underlying competency (in this case scientific literacy) which is regarded as a latent variable that cannot be measured directly (Moosbrugger, 2007; Rost, 2004). To analyze the NEPS data, the one-parameter Rasch model (Rost, 2004) was used. One of the important features of the Rasch model is that it creates a continuum on which both student performance and item difficulty can be located. Low performers and easy items are located at one end of the continuum while high performers and difficult items are located at the opposite end of the continuum.

3.2.1 Estimating the person parameters – The weighted likelihood estimator (WLE)

The programme *ConQuest* (Wu, Adams, Wilson, & Haldane, 2007) was used to analyze the data and to investigate the psychometric quality of the tests (e.g., to compute item difficulties, test scores, differential item functioning and dimensional analyses). The data were scaled using a one-parameter Rasch model and Marginal Maximum Likelihood (MML) techniques for parameter estimation. Weighted Likelihood Estimates (WLE) were used as person parameters (ability estimates).

The MTF items were analyzed using the partial credit model. Each of the four true-false answers was given 0.5 points if it was answered correctly. Thus, the maximum score of a MTF item was two points. Three different codes existed for miss-

ing values: omitted, invalid and not reached. In the analyses, however, all these codes were treated as "missing".

3.2.2 Item quality

In order to create a test that ultimately facilitates a reliable and valid measurement of scientific literacy in the main study, item quality has to be evaluated. At the same time, the structure of the NEPS science framework has to be considered during the item selection process.

Complying with international standards, NEPS focuses on four values when judging item quality: item discrimination, fit between persons' abilities and item difficulties, *t* values (and mean squares) and differential item functioning (DIF).

Item discrimination[1] in the context of item selection relates to the correlation between an item *i* and the result of the test *t* (Rost, 2004). Items remain in the test when their discrimination reaches values > .30. This is even stricter than the criterion in PISA, for example, where items are flagged as "dodgy" if their discrimination value is lower than .20 (OECD, 2009).

With respect to *item difficulty,* the aim in NEPS is to create a test in which the distribution of item difficulties matches the distribution of the target persons' abilities. Due to Rasch scaling, item difficulties and person abilities are located on the same scale. Thus, items that are too easy or too difficult for the test sample can be identified and eliminated from the test. Fixing the mean person ability at a value of zero means that the mean item difficulty should also be as close to zero as possible.

The third criterion is the infit of the items. The test statistics that are considered here are the *t value* and the *mean square* (MNSQ, weighted fit). In ConQuest analyses, the *t* value serves as an indicator for a violation of the Rasch criterion of item homogeneity. *t* values should lie between 2 and -2 (Wright, Mead, & Bell, 1980; Smith, 1995). The MNSQ accounts for the fact that the relation between *t* value and item discrimination is influenced by the sample size. If the mean square has a value of 1, the observed value equals the expected one. Values > 1 signalize an unexpectedly high variance and values < 1 an unexpectedly low variance in the deviation of the item's discrimination (Wilson, 2005). In NEPS, items showing a mean square within the range of 0.85 to 1.15 are regarded as acceptable which is well within internationally accepted standards (Adams & Khoo, 1996).

Finally, analyses of DIF are carried out in order to identify items showing different statistical properties for certain subgroups of the sample (the compared test persons being equally competent). These items have to be removed from the test. Typically, DIF analyses are carried out for gender, migration background and school type. In the NEPS pilot studies, DIF analyses could only be carried out for gender because there was no oversampling of persons with a migration background and the sample sizes per school type were also limited.

1 Items with good discrimination values distinguish well between persons with different abilities.

Concerning gender DIF, items with differences in item difficulty exceeding 0.3 were regarded as problematic (OECD, 2009) and eliminated from the test. In some cases, items with higher DIF values had to remain in the test due to constraints caused by the unit structure or the context-component combinations. In these cases, it was ensured that the mean DIF of the complete test was close to zero.

3.2.3 Psychometric properties – Checking for reliability and validity aspects

3.2.3.1 Reliability. Reliability is one of the important criteria when checking for test quality. It refers to the accuracy and consistency of a test instrument's measurement. The reliability of a test increases when the amount of random measurement errors decreases (Schermelleh-Engel & Werner, 2007). By using the ConQuest software, the probabilistic reliability of the NEPS science tests is computed with the Andrich method which accurately estimates the reliabilities of tests containing more than 20 items. Tests measuring achievement should reach reliability values higher than .70 (Schermelleh-Engel & Werner, 2007).

3.2.3.2 Internal validity – Dimensional analyses. In order to check the internal validity, dimensional analyses are carried out to find out whether the theoretically postulated competency structure applies to the data. Due to the short testing time and the limited number of items, it is not possible to get valid scores for each of the content and process related components described in Section 2.1. As a consequence, a one dimensional model is used for computing one WLE score for every test person as an indicator of his or her scientific literacy. In order to verify whether this is a valid way of assessing a person's scientific literacy, this model is compared (concerning its fit) with the results from a two-dimensional model separating the content and process related components *knowledge of science* and *knowledge about science*. For comparison, the Bayes information criterion (BIC) is computed for both models. The model showing the lower BIC value is the one that fits the data best.

3.2.3.3 External validity – Correlations with external measures. Checking the external validity of a test means correlating the test results with external measures that relate more or less closely to the construct of scientific literacy.

Due to limited testing time, only first steps could be taken concerning the validation of the NEPS science test. Two different scales were included in the test or the student questionnaires as first validation tools.

In kindergarten, children's interest in science is measured by a scale that was originally developed for the *Study on Competence Development in Elementary Science Education* (SNaKE; Lankes, Steffensky, & Carstensen, 2009). It was adapted for use in NEPS and measures the interest in science, music, art and reading. Each of these interests is assessed on a three item base. Corresponding to the as-

sessment of scientific literacy in kindergarten, children's interests are also measured in a picture-based way. The NEPS children are asked, for example, to imagine their birthday and then choose from four pictures what they would like first, second, third and fourth best for a present (see Figure B1).

In Grades 6 and 9, a slightly adapted version of the PISA science activities scale SCIEACT (Frey et al., 2009) is used to assess students' interest in science by asking them about their out-of-school science activities (see Table B1).

When correlating students' achievement in the science tests with the interest scales mentioned in this section, the *discriminant validity* of the test is checked. Since two different constructs are measured, an achievement and an interest scale, the correlation between these scales should be lower than the correlation between two achievement scales but should reach values between $r = .20$ and $r = .30$ (OECD, 2007; Krapp, Schiefele, & Schreyer, 1993). However, these correlations should still be significant according to theories concerning interest and epistemic orientation (Prenzel, 1988; Krapp, 1996, Krapp & Prenzel, 2011).

4. Results

The following section gives a short overview of the results based on pilot study data from kindergarten, Grade 6 and 9.

4.1 Item Quality

4.1.1 Item discrimination

In NEPS, items with discrimination values < .30 are eliminated from the test. In the pilot studies, 12 of the 47 kindergarten items, 18 of the 49 Grade 6 items and 9 of the 45 Grade 9 items were below the item discrimination threshold of .30 and were thus removed from the test.

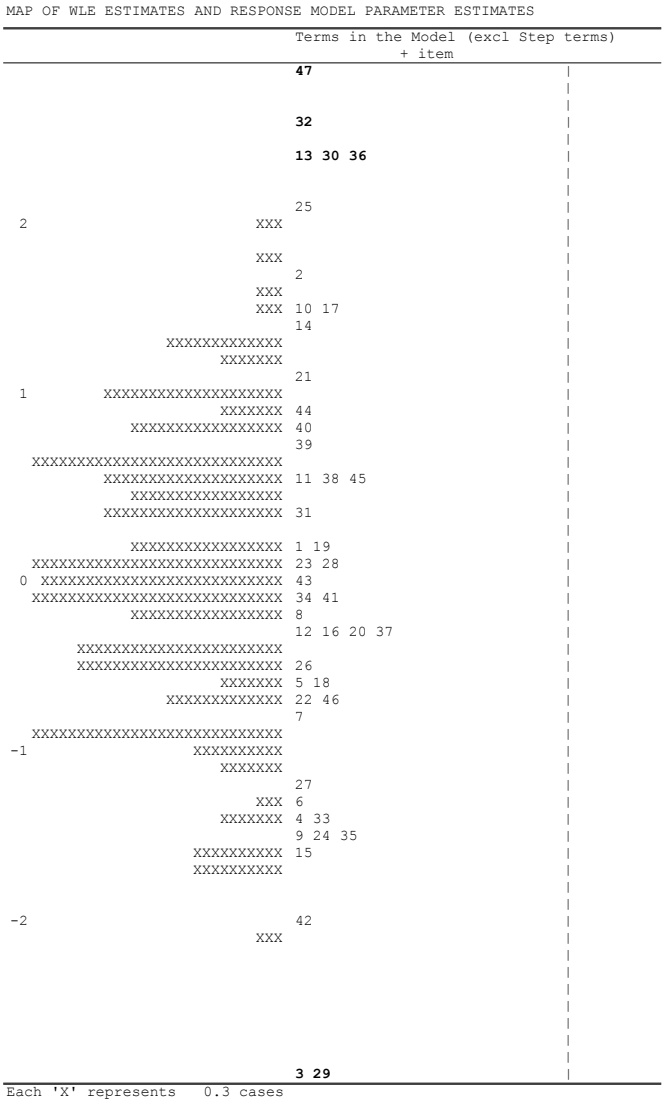
4.1.2 Item difficulties and person abilities

The aim of NEPS is to create a test that makes it possible to assess test persons' scientific literacy in all areas of the ability continuum. Consequently, the distribution of item difficulties should match the distribution of the test persons' abilities. As an example, Figure 3 illustrates the results of the kindergarten cohort: On the left, the distribution of person abilities is presented and on the right, the distribution of item difficulties is presented. The mean person ability is set at zero. The mean item difficulty in kindergarten is 0.04 and the distribution of item difficulties fits the distribution of person abilities rather well. Items 13, 30, 36, 32, 47, 3, 29

were removed from the test as their difficulties were outside the range of the test persons' abilities.

In Grade 6, the mean difficulty of the piloted items was 0.15. The test was regarded as slightly too difficult with regard to the Wright-Map of Grade 6 and some of the difficult items were removed. In Grade 9, the mean item difficulty of the piloted items amounted to -0.29. Hence, the test was too easy for that age group and the Wright-Map showed a lack of difficult items. Based on existing and rather difficult items, an additional unit was developed and included in the test.

Figure 3: Wright-Map of the NEPS kindergarten test results (science test)



4.1.3 *t* values (MNSQ)

If the *t* value of an item fell within a range of 2.0 to -2.0, it was included in the test, while the weighted MNSQ value had to be between 0.85 and 1.15.

The results show that 44 of the 47 kindergarten items met these requirements. Two items exceeded the threshold of 2.0 with values of 2.8 (weighted MNSQ = 1.17) and 4.7 (weighted MNSQ = 1.28) and one item had a *t* value < -2.0 (-2.6) with a weighted MNSQ of 0.85.

In Grade 6, 36 of the 49 items fit in the postulated *t*-value range. Six items exceeded a *t* value of 2.0 (2.3 to 3.3) with weighted MNSQ values ranging from 1.09 to 1.13 and seven items showed values < -2.0 (-2.1 to -4.2) with weighted MNSQs ranging from 0.92 to 0.85.

For Grade 9, 43 of the 45 items fit in the range that was set for the *t* values. Two items exceeded the range (*t* value = 2.6, weighted MNSQ = 1.14 and *t* value = 3.2, weighted MNSQ = 1.17).

Items that did not meet the criteria for the *t* and MNSQ values were removed from the test. If only the *t* value of an item was not within the defined range, the item was only excluded from the test if the selection criteria for item difficulty or item discrimination were not met either.

4.1.4 DIF analyses

Analyses of gender differences illustrate that some items are easier for male test persons, whereas others are easier for female test persons. Therefore, DIF values have to be taken into account during the item selection processes in order to create a final test with minimal differential item functioning concerning gender. For all of the NEPS pilot tests, mean gender DIFs were close to zero (kindergarten: -0.049; Grade 6: 0.001; Grade 9: -0.004).

4.2 Reliability and validity

In kindergarten, the WLE reliability was .81, in Grade 6, .84 and in Grade 9, .83. All reliabilities can thus be regarded as being more than sufficient (Schermelleh-Engel & Werner, 2007).

Results from assessing the dimensionality of the NEPS science tests for kindergarten, Grade 6 and Grade 9 are presented in Table 3.

Table 3: Dimensional analyses for the tests in kindergarten, Grade 6 and Grade 9

Age cohort	BIC 1dim	BIC 2dim	Favoured model
Kindergarten	6,813.94	6,821.09	1dim
Grade 6	21,117.52	21,105.97	2dim
Grade 9	9,893.86	9,893.45	2dim

Note. BIC = Bayesian Information Criterion; 1dim = one-dimensional; 2dim = two-dimensional.

In Grade 6 and Grade 9, the two-dimensional model represents the data slightly better than the one-dimensional model. For both grades, however, the two sub-dimensions are highly correlated (r = .98. in Grade 6 and r = .97 in Grade 9). Hence, it is acceptable to consider the test as being one-dimensional.

For kindergarten, scientific literacy was correlated with the children's interest in science, music, art and reading. First, these interest scales were analyzed in terms of their reliabilities. Initially, each interest scale contained three items, but in order to optimize the homogeneity of the scales according to results from the reliability analyses, one item had to be deleted from each scale. However, the Cronbach's alpha[2] values of the interest scales remained very low (Cronbach's alpha < .38). Although the correlation between scientific literacy and interest in science is significant (r = .19[3], p < .05) and exceeds the correlations between scientific literacy and interest in music (r = -.15, p < .05), art (r = -.01, n.s.) and reading (r = .13, n.s.), all these correlations should be considered as a first approach to validation, due to the low reliabilities of the interest scales.

In Grade 6 and Grade 9, correlations were computed between scientific literacy and an adaptation of the PISA science activities scale (Frey et al., 2009, see Table B1). Its WLE reliabilities were .65 (Grade 6) and .59 (Grade 9). In Grade 6, the correlation between scientific literacy and the science activities scale was not significant (r = .02). In Grade 9, the correlation was significant but low (r = .18, p < .05).

5. Discussion and Outlook

The aim of presenting the NEPS science framework and first results from pilot studies of different age cohorts was to illustrate the theoretical background, the item development and the item selection process for the final NEPS science tests. The presented results demonstrate the importance of pilot studies in test development processes. A reliable test can only be constructed with well-designed pilot

2 Cronbach's alpha is a coefficient that rates the internal consistency or the correlation of the items in a test (values lie between 0 and 1). Since different aspects of a construct should be measured, a scale showing strong internal consistency should reach a moderate correlation between items (.70 to .90).

3 Due to different scale level (the interest scale being a ranking scale), Spearman correlations were computed between the interest scales and science competency (scientific literacy?).

studies and strict item selection criteria for the final test. In order to obtain a sufficient number of items fulfilling these criteria, three times more items had to be developed and piloted than are eventually needed for the final test.

All in all, the first research question of whether the statistical properties of the items comply with internationally accepted standards can be answered positively. The quality of the tests that entered the NEPS pilot studies was already promising. Most of the items showed statistical properties that comply with the demands of internationally accepted standards. Items that did not reach a discrimination > .30, exceeded the extremes of persons' abilities or did not fulfill the Rasch criterion of item homogeneity were removed from the test. There is evidence for gender DIF in single items. However, none of the tests' DIF means were striking.

For the second research question, which addressed the reliability and validity of the tests, first evidence could be found that the tests measure scientific literacy in a reliable way. WLE reliabilities of the tests in kindergarten, Grade 6 and Grade 9 reached values > .80 and can thus be considered as being more than sufficient for a reliable test (Schermelleh-Engel & Werner, 2007).

The other important criterion concerning the quality of a test is its validity. The first approach was to confirm the tests' internal validity. The results either show that the one-dimensional model fits the data best (kindergarten) or they were slightly in favor of a two-dimensional model with two highly correlated sub-dimensions (Grade 9). These results are not surprising as the science test in fact measures two different content areas – KOS and KAS. Results from PISA 2006 have also shown that these sub-dimensions are highly correlated (Prenzel et al., 2007). Nonetheless, the aim of the final NEPS tests is to assess scientific literacy consistently using the same model for all age cohorts. As the data we presented in this paper originate from pilot studies (before item selection), it remains to be seen whether the final tests of the main studies are one-dimensional.

The first approaches towards checking the tests' external validity by correlating scientific literacy with interest in science have to be interpreted with caution. The correlations between scientific literacy and the interest in science scale (kindergarten) and the science activities scale (Grade 9) are significant but low and are comparable to values from PISA (OECD, 2007) while the correlation between scientific literacy and the science activities scale in Grade 6 is not significant. This result might be due to the fact that the science activities scale was originally constructed for Grade 9 children and might not be valid for younger children. The correlations between scientific literacy and interest in science in kindergarten exceed the correlations between scientific literacy and interest in music, art or reading. However, due to low reliability values, these results can only be seen as a first indication. As testing time in NEPS is strictly limited, it will be impossible to place more scales on science activities or interest in science in the NEPS assessments. In order to evaluate the external validity of the science tests, additional studies will have to be set up.

The key aspect of a test is its construct validity. In order to confirm construct validity, a study has to demonstrate that the test scores and the prediction of a the-

oretical trait – in this case scientific literacy – are connected. Therefore, the NEPS science tests need to be validated by using existing measures of scientific literacy. As this cannot be accomplished within the NEPS schedule, additional studies will have to be carried out. One way to validate the NEPS Grade 9 science test was to use the assessment period of PISA 2012 to combine the PISA test, the test of the German educational standards (which was administered at the same time) and the NEPS science test to measure students' scientific literacy and compare the results of the three tests. This validation study was carried out in early summer 2012.

Another area of research is to link the science tests for different age groups in order to ensure that the NEPS science tests measure the same construct in every age group. The studies have not yet been linked because the intervals between the age groups (e.g., kindergarten, Grade 6 and Grade 9) were too large to permit an appropriate link. The larger the interval, the more random error variance is likely to influence the measurement. Hence, intervals between two samples that are to be linked should be as short as possible but long enough to detect differences in scientific literacy. The first linking studies will be possible in 2013 when the kindergarten science test can be linked to the test in Grade 1, and the Grade 9 test can be linked to the test in Grade 11.

First results from the main studies using the final NEPS science tests for kindergarten and Grade 9 were released as public use files in the second half of 2012. The data of the main study in Grade 6 will be available in 2013. In the following years, a clearer picture should emerge as to how scientific literacy develops over the lifespan. Questions that can be answered with the main study data concern, for example, the influence of competencies on decisions at critical transitions in life or the extent to which the development of competencies is influenced by family background, peer groups or the arrangement of learning opportunities in formal and informal learning environments.

References

AAAS – American Association for the Advancement of Science. (1993). *Benchmarks for science literacy. Project 2061*. New York, NY: Oxford University Press.

AAAS – American Association for the Advancement of Science. (2007). *Atlas of science literacy*. Washington, DC: AAAS.

AAAS – American Association for the Advancement of Science. (2009). *Benchmarks for science literacy. Project 206*. Retrieved from http://www.project2061.org/publications/bsl/online/index.php

Adams, R., & Khoo, S. (1996). *Quest: The interactive test analysis system – version 2.1* (Technical report). Camberwell, Australia: Australian Council for Educational Research.

Ahlgren, A., & Rutherford, F. J. (1991). *Science for all Americans*. New York, NY: Oxford University Press.

Albanese, M. A., & Sabers, D. L. (1988). Multiple true-false items: A study of interitem correlations, scoring alternatives, and reliability estimation. *Journal of Educational Measurement, 25*, 111–123.

Bybee, R. W. (1997a). Towards an understanding of scientific literacy. In W. Gräber & C. Bolte (Eds.), *Scientific literacy – An international symposium* (pp. 37–68). Kiel, Germany: Institut für die Pädagogik der Naturwissenschaften (IPN).

Bybee, R. W. (1997b). *Achieving scientific literacy: From purposes to practices*. Portsmouth, NH: Heinemann Educational Books.

Bybee, R. W., McCrae, B. J., & Laurie, R. (2009). PISA 2006: An assessment of scientific literacy. *Journal of Research in Science Teaching, 46*, 865–883.

Bybee, R. W., & *PISA 2006 Science Expert Group*. (2009). An assessment framework for scientific literacy. In R. W. Bybee & B. J. McCrae (Eds.), *PISA science 2006. Implications for science teachers and teaching* (pp. 15–26). Arlington, VA: NSTA press.

Carey, S. (1991) *The nature of science interview*. Unpublished manuscript.

Chinn, C. A., & Brewer, W. F. (1998). An empirical test of a taxonomy of responses to anomalous data in science. *Journal of Research in Science Teaching, 35*, 623–654.

Council of Ministers in Education. (1997). *Common framework of science learning outcomes*. Toronto, Canada: Council of Ministers of Education.

European Commission. (2007). *Key competences for lifelong learning. European reference framework*. Luxembourg, Luxembourg: Office for Official Publications of the European Communities. Retrieved from: http://ec.europa.eu/dgs/education_culture/publ/pdf/ll-learning/keycomp_en.pdf

Fensham, P. J. (2009). Real world contexts in PISA science: Implications for context-based science education. *Journal of Research in Science Teaching, 46* (8), 884–896.

Frey, A., Taskinen, P., Schütte, K., Prenzel, M., Artelt, C., Baumert, J., Blum, W., Hammann, M., Klieme, E., & Pekrun, R. (Eds.). (2009). *PISA 2006 Skalenhandbuch – Dokumentation der Erhebungsinstrumente*. Münster, Germany: Waxmann.

Frisbie, D. A. (1992). The status of multiple true-false testing. *Educational Measurement: Issues and Practices, 5*, 21–26.

Gräber, W., Nentwig, P., Koballa, T., & Evans, R. (Eds.). (2002). *Scientific Literacy. Der Beitrag der Naturwissenschaften zur Allgemeinen Bildung*. Opladen, Germany: Leske + Budrich.

Haladyna, T. M. (1992). The effectiveness of several multiple-choice formats. *Applied Measurement in Education, 5*, 73–88.

Haladyna, T. M., & Downing, S. M. (1989). The validity of a taxonomy of multiple-choice item-writing rules. *Applied Measurement in Education, 1*, 51–78.

Hodson, D. (1992). In search of a meaningful relationship: An exploration of some issues relating to integration in science and science education. *International Journal of Science Education, 14*(5), 541–562.

Hurd, P. D. (1998). Scientific literacy: New minds for a changing world. *Science Education, 82*, 407–416.

Klahr, D. (2000). *Exploring science*. Cambridge, MA: MIT Press.

Klahr, D., & Dunbar, K. (1988). Dual space search during scientific reasoning. *Cognitive Science, 12*, 1–55.

Klieme, E., Avenarius, H., Blum, W., Döbrich, P., Gruber, H., Prenzel, M., Reiss, K., Riquarts, K., Rost, J., Tenorth, H.-E., & Vollmer, H. J. (2004). *The development of national education standards. An expertise* (BMBF Education Reform, Vol. 1). Bonn, Germany: BMBF.

Klieme, E., & Leutner, D. (2006). Kompetenzmodelle zur Erfassung individueller Lernergebnisse und zur Bilanzierung von Bildungsprozessen. *Zeitschrift für Pädagogik, 52*(6), 876–903.

KMK – Kultusministerkonferenz. (2005a). *Beschlüsse der Kultusministerkonferenz: Bildungsstandards im Fach Biologie für den Mittleren Schulabschluss. Beschluss vom 16.12.2004*. München, Germany: Luchterhand.

KMK – Kultusministerkonferenz. (2005b). *Beschlüsse der Kultusministerkonferenz: Bildungsstandards im Fach Chemie für den Mittleren Schulabschluss. Beschluss vom 16.12.2004*. München, Germany: Luchterhand.

KMK – Kultusministerkonferenz. (2005c). *Beschlüsse der Kultusministerkonferenz: Bildungsstandards im Fach Physik für den Mittleren Schulabschluss. Beschluss vom 16.12.2004*. München, Germany: Luchterhand.

Koslowski, B. (1996). *Theory and evidence. The development of scientific reasoning*. Cambridge, MA: MIT Press.

Krapp, A. (1996). Die Bedeutung von Interesse und intrinsischer Motivation für den Erfolg und die Steuerung schulischen Lernens. Theorie und Praxis der Unterrichtsforschung. In G. W. Schnaitmann (Ed.), *Theorie und Praxis der Unterrichtsforschung* (pp. 88–111). Donauwörth, Germany: Auer.

Krapp, A., & Prenzel, M. (2011). Research on interest in science: Theories, methods, and findings. *International Journal of Science Education, 33*(1), 27–50.

Krapp, A., Schiefele, U., & Schreyer, I. (1993). Metaanalyse des Zusammenhangs von Interesse und schulischer Leistung. *Zeitschrift für Entwicklungspsychologie und Pädagogische Psychologie, 10* (2), 120–148.

Lankes, E.-M., Steffensky, M., & Carstensen, C. H. (2009). *Studie zur naturwissenschaftlichen Kompetenzentwicklung im Elementarbereich (SNaKE)*. Retrieved from http://www.uni-bamberg.de/fileadmin/psymethodenbf/homepage_snake_4_seiten.pdf.

Laugksch, R. C. (2000). Scientific literacy: A conceptual overview. *Science Education, 84* (1), 71–94.

Linacre, J. K. (1994). Sample size and item calibration stability. *Rasch Measurement Transactions,* 7 (4), 328.

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

Masnick, A. M., & Klahr, D. (2003). Error matters: An initial exploration of elementary school children's understanding of experimental error. *Journal of Cognition and Development, 4*, 67–98.

Miller, J. D. (1983). Scientific literacy: A conceptual and empirical review. *Daedalus, 112* (2), 29–48.

Millar, R., & Osborne, J. F. (Eds.). (1998). *Beyond 2000: Science education for the future*. London, United Kingdom: King's College London.

Moosbrugger, H. (2007). Item-Response-Theorie. In H. Moosbrugger & A. Kelava (Eds.), *Testtheorie und Fragebogenkonstruktion* (pp. 99–112). Heidelberg, Germany: Springer.

National Research Council. (1998). *National science education standards*. Washington, DC: National Academy Press.

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

OECD – Organisation for Economic Co-operation and Development. (2007). *PISA 2006 Volume 2: Data/Données*. Paris, France: OECD.

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

Osborne, J. F., & Dillon, J. (2008). *Science education in Europe: Critical reflections. A report to the Nuffield Foundation*. London, United Kingdom: King's College London.

Prenzel, M. (1988). *Die Wirkungsweise von Interesse. Ein Erklärungsversuch aus pädagogischer Sicht*. Opladen, Germany: Westdeutscher Verlag.

Prenzel, M. (2000). Lernen über die Lebensspanne aus einer domänenspezifischen Perspektive: Naturwissenschaften als Beispiel. In F. Achtenhagen & W. Lempert

(Eds.), *Lebenslanges Lernen im Beruf – seine Grundlegung im Kindes- und Jugendalter. Band IV. Formen und Inhalte von Lernprozessen* (pp. 175–192). Opladen, Germany: Leske + Budrich.

Prenzel, M., Schöps, K., Rönnebeck, S., Senkbeil, M., Walter, O., Carstensen, C., & Hammann, M. (2007). Naturwissenschaftliche Kompetenz im internationalen Vergleich. In M. Prenzel, C. Artelt, J. Baumert, W. Blum, M. Hammann, E. Klieme, & R. Pekrun (Eds.), *PISA 2006. Die Ergebnisse der dritten internationalen Vergleichsstudie* (pp. 63–105). Münster, Germany: Waxmann.

Prenzel, M., & Seidel, T. (2008). Erwerb naturwissenschaftlicher Kompetenzen. In W. Schneider & M. Hasselhorn (Eds.), *Handbuch Pädagogische Psychologie* (pp. 608–618). Göttingen, Germany: Hogrefe.

Rost, J. (2004). *Lehrbuch Testtheorie – Testkonstruktion*. Bern, Switzerland: Hans Huber.

Schermelleh-Engel, K., & Werner, C. (2007). Methoden der Reliabilitätsbestimmung. In H. Moosbrugger & A. Kelava (Eds.), *Testtheorie und Fragebogenkonstruktion* (pp. 113–133). Heidelberg, Germany: Springer.

Senatsverwaltung für Bildung, Jugend und Sport Berlin. (2006). *Rahmenlehrplan für die Sekundarstufe I, Jahrgangsstufe 7-10. Hauptschule, Realschule, Gesamtschule, Gymnasium. Physik*. Retrieved from: https://www.berlin.de/imperia/md/content/senbildung/schulorganisation/lehrplaene/sek1_physik.pdf?start&ts=1150101938&file=sek1_physik.pdf

Shwartz, Y., Ben-Zvi, R., & Hofstein, A. (2006). The use of scientific literacy taxonomy for asssessing the development of chemical literacy among high-school students. *Chemistry Education Research and Practice, 7*(4), 203–225.

Smith, C. L., Wiser, M., Anderson, C. W., & Krajcik, J., (2006). Implications on research on children's learning for standards and assessment: A proposed learning for matter and the Atomic Molecular Theory. *Measurement: Interdisciplinary Research & Perspective, 4*(1), 1–98.

Smith, R. M. (1995). *Using item mean squares to evaluate fit to the Rasch model*. Paper presented at the Annual Meeting of the American Educational Research Association. San Francisco, CA.

Trefil, J. (2008). *Why science?* Arlington, VA: NSTA Press.

UNESCO – United Nations Educational, Scientific and Cultural Organization. (1993). *International forum on scientific and technological literacy for all. Final report*. Paris, France: UNESCO.

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

Wilson, M. (2005). *Constructing measures: An item response modelling approach*. Mahwah, NJ: Lawrence Erlbaum Associates.

Wright, B. D., Mead, R. J., & Bell, S. R. (1980). *BICAL: Calibrating items with the Rasch Model*. Chicago, IL: Research Memorandum.

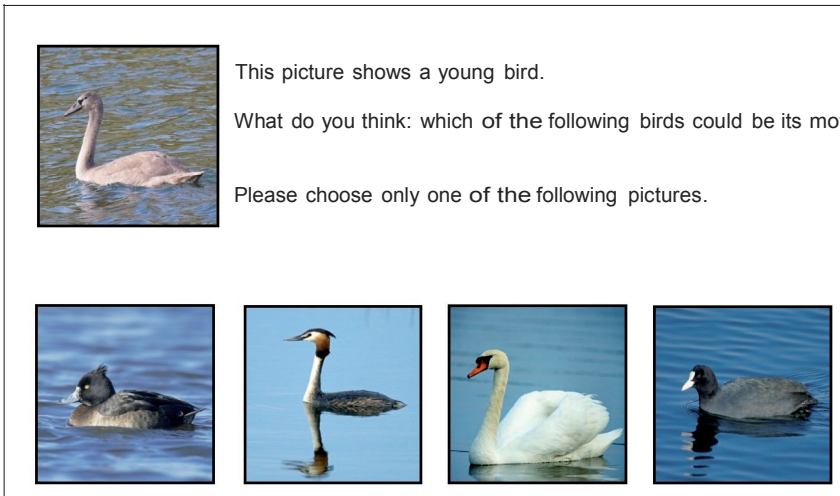
Wu, M. L., Adams, R. J., Wilson, M. R., & Haldane, S. A. (2007). *ACERConQuest Version 2: Generalised item response modelling software*. Camberwell, Australia: Australian Council for Educational Research.

Zimmerman, C. (2007). The development of scientific thinking skills in elementary and middle school. *Developmental Review, 27*(2), 172–223.

Appendix A

Item examples of the NEPS test measuring scientific literacy[4]

Figure A1: Example of an item measuring scientific literacy in kindergarten[5] (context: environment, component: development)



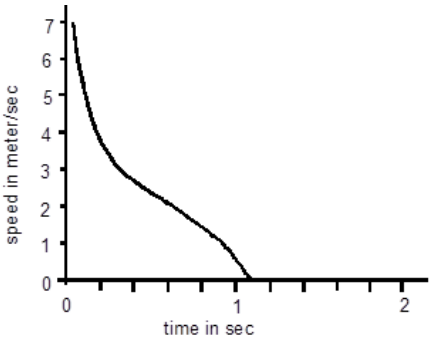
4 Please note that the presented items for measuring scientific literacy are not part of the main test. They had to be eliminated from the test after the pilot studies due to a lack of item quality. They only serve as examples of the item structure.

5 This item is part of a picture puzzle with which the children are playing during the summer party in kindergarten.

Figure A2: Example of an item measuring scientific literacy in Grade 6 (context: technology, component: scientific enquiry and scientific reasoning)

The flight of a shuttlecock

This figure shows how the speed of a shuttlecock changes after it has been hit by a racket.





Which one of the following statements can be deduced from this figure? Please take your time to look at the figure in detail!

Please tick the box in front of the right answer! Please tick only one box!

☐	After a few meters the shuttlecock falls to the ground.
☐	The shuttlecock starts to trundle.
☐	The speed of the shuttlecock decreases rapidly.
☐	The shuttlecock flies in a curve.

Appendix B

Examples of items measuring interest in science and science activities

Figure B1: Example of an item measuring children's interest in kindergarten

Imagine it is your birthday. I am going to show you four pictures of things you could wish for. Which of following things would you like best for your birthday?

A: A box containing children´s books



B: A box containing things for doing handicrafts



C: A box containing music CDs

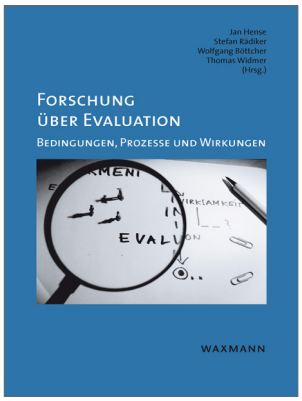


D: A box containing an experimental kit



Table B1: The adapted PISA science activities scale (Frey et al., 2009)

How often do you do these things?	never	rarely	some-times	often
a) Watch TV programmes about broad science	☐	☐	☐	☐
b) Borrow or buy books on broad science topics	☐	☐	☐	☐
c) Visit websites about broad science topics	☐	☐	☐	☐
d) Read broad science magazines or science articles in newspapers	☐	☐	☐	☐
e) Attend a science club	☐	☐	☐	☐



Jan Hense, Stefan Rädiker,
Wolfgang Böttcher, Thomas Widmer (Hrsg.)

Forschung über Evaluation

Bedingungen, Prozesse und Wirkungen

2013, 282 Seiten, br., 29,90 €
ISBN 978-3-8309-2829-4
E-Book-Preis: 26,99 €

Dieser Band stellt eine transdisziplinäre Plattform für Forschungsprojekte bereit, die Evaluationen und Evaluationssysteme als Ganzes oder in Teilaspekten zum Gegenstand haben. Mittels vielfältiger theoretischer und methodischer Zugänge untersuchen die Autorinnen und Autoren Evaluationsaktivitäten in Anwendungsfeldern wie Schule, Hochschule, Weiterbildung, Mentoring, Sozialer Arbeit oder Entwicklungszusammenarbeit. Vertreten sind dabei verschiedene Fachdisziplinen wie Pädagogik, Psychologie, Politikwissenschaft oder Soziologie.

WAXMANN
Münster · New York · München · Berlin