KNUT NEUMANN, HANS E. FISCHER and ALEXANDER KAUERTZ

FROM PISA TO EDUCATIONAL STANDARDS: THE IMPACT OF LARGE-SCALE ASSESSMENTS ON SCIENCE EDUCATION IN GERMANY

Received: 14 September 2009; Accepted: 8 February 2010

ABSTRACT. The German education system does not traditionally rely on standardized testing. However, when the Programme for International Student Assessment (PISA) study revealed an average performance of German students compared to other participating countries, a particular proportion of low-performing students, and remarkable disparities between the federal states, German policy makers decided for a major reform of the education system. A core piece of this reform was the introduction of National Education Standards. For science education, these standards were heavily influenced by the PISA results and its underlying framework. That is, with the standards, a paradigm shift took place from the German notion of *Bildung* towards the Anglo-American notion of *literacy*. With the introduction of these standards, a new field of empirical educational research was created: research on models of scientific literacy or competency models as a basis of benchmarking the standards. This article describes the German education system before PISA, summarizes the major findings from PISA, and describes how these findings informed the formulation of the performance standards for science education. It also details the measures undertaken to benchmark these standards. Finally, it provides insight into the issues with developing and benchmarking performance standards and points out future areas of research on evidence-based decision making in educational policy.

KEY WORDS: competencies, Germany, PISA, science education, standards

Large-scale assessments do not have a long tradition in Germany. Current practices started with German participation in the First International Science Study from 1962 to 1972, which had no particular impact on German education. The second international comparison in which Germany took part was the Third International Mathematics and Science Study (TIMSS) in 1995. The reason for the long break lies in a particular paradigm underpinning the German education system. Within this paradigm, called *Humanistische Bildung*, the focus is on the self-formation of the individual. Roughly speaking, it is believed that individuals carry the desire to develop within themselves and that this desire needs to be nurtured (Sorkin, 1983). Consequently, education standards and respective assessments were not considered meaningful. However, with a growing need for competitiveness in a global economy and the mediocre performance of Germany in the 1995 TIMSS study,

International Journal of Science and Mathematics Education (2010) 8: 545–563

German policy makers finally wanted to be informed about German students' achievement in comparison to that of other leaders in the global economy (Ständige Konferenz der Kultusminister der Länder der Bundesrepublik Deutschland [Standing Conference of the Ministers of Education and Cultural Affairs of the Länder in the Federal Republic of Germany], 1997).

The findings of the 1995 TIMSS study came as a shock to German teachers, researchers, and policy makers. German students' science and mathematics achievement was revealed as barely average—devastating results not in line with the firm belief in the superiority of the German education system. Even worse was that the results were confirmed by the Programme for International Student Assessment (PISA). German students' achievement in reading, mathematics, and science in PISA 2000 was found to be average at best; as well, a number of students was found to fail in achieving even the lowest level of literacy. Based on these findings, policy makers took measures—programs to improve instructional quality were created, funding of schools was increased, and most importantly, a reform of the education system itself was initiated—to address the deficits of German science education revealed by PISA.

The core of the reform was a growing acceptance of the notion of literacy over the traditional German notion of education (see, e.g. Klieme, Avenarius, Blum, Döbrich, Gruber, Prenzel et al., 2003). Based on this paradigm shift, the National Education Standards (NES) were introduced for the science subjects of biology, chemistry, and physics. Informed by the PISA findings, the NES describe scientific competencies students are expected to have acquired at the end of their lower secondary level education (i.e. about 16 years old). This is a particular change to an education system—because a policy in which science learning is assessed by an external authority affects practice differently than does a policy in which the school and teachers are required to provide and justify the summative assessment internally (Fensham, 2009b). The reform of the German education system, in particular the introduction of the NES for the science subjects, seems to be an interesting object of research on how large-scale assessments initiated and informed educational reforms.

This article presents a historical summary of the reform of the German education system using the example of science education. To provide insight into the specifics of science education, we first give a brief introduction to the German education system prior to PISA. Subsequently, the most important findings of the first cycle of PISA from 2000 to 2006 will be summarized based on the international and national reports. We will then elucidate how the PISA results informed a fundamental reform of the German education system and the problems that occurred during this process. We will focus in particular on the genesis of the NES for the lower secondary school level and how this particular change in the educational system effectively strengthened empirical research in the field of science education. Finally, we will discuss if and how the described measures may have contributed to improving educational quality in general and quality of instruction in particular. We will use this discussion to establish future areas for research on the effects of evidence-based policy making in education.

Science Education in Germany before PISA

Education and the particular vision of education are culturally specific. As such, education reforms have to be judged based on the particular cultural aspects of an education system. The noteworthiness of the reform of German science education triggered by the PISA results can only be understood in light of the peculiarities of the German education system and the role of science education within this system.

The German Education System

The history of the German education system and the fact that Germany is a confederation of individual states have led to a particularly complex education system, a detailed description of which is beyond the scope of this article (for a detailed description, see Döbert, 2007). However, the characteristics of the education system prior to PISA are of particular relevance for understanding the German discussion of the PISA findings: the specific German notion of education, the fact that the authority over the education system is held by the German federal states, and a system that heavily relies on tracking.

Bildung. Although the traditional German term *Bildung* is often translated as *education*, the notion of Bildung differs substantially from the Anglo-American notion of education. It is particularly influenced by the vision of *Humanistische Bildung*, which focuses on the development of the individual; that is, the perfection of the personality of an individual. Klafki (1996) pointed out that not only functional abilities of an individual have to be considered when organizing education but also abilities related to individual formation, the so-called conception of man, with only indirect connections to social matter. As such, the esthetic, ethical, and intellectual development of the human being was at the center of attention, whereas economic and social considerations were of no particular relevance (for a more detailed description of Bildung, see Fischer, Kauertz & Neumann, 2008). As the focus was on the development of the individual, general education standards and the assessment of student achievement with respect to these standards were not part of the educational policy.

The authority for what is taught and how it is taught in the German education system was not a central curriculum and assessment authority but was, within limits, in the hands of teachers (Fensham, 2009b). The education system defined the general objectives of education based on which written curricula described the contents to be covered in a particular subject, but teachers were free to set the educational agenda within these limits. Each teacher was responsible for the pedagogy and, consequently, for the assessment of students' learning (Fensham). Thus, the education system prior to PISA was an input-oriented education system in which policy makers exerted control on curricula, teacher education, school funding, etc. In contrast, the Anglo-American way of defining standards and assessing student achievement with respect to these standards would be considered as an output-oriented system (Klieme et al., 2003).

Federalism. Based on the Basic Law of the Federal Republic of Germany and an agreement between the federal states and the federal government, each federal state holds the authority over its education system. As Germany consists of 16 federal states, one may expect this to lead to 16 different educational systems. However, as the Basic Law also guarantees equality of opportunity for all citizens, similar educational systems emerged (Döbert, 2007): students first go to a common elementary school, after which they follow one of several school tracks; only if they successfully complete the highest-level track are they allowed to go to university.

Still, there are some differences. Science, for example, would be taught in most federal states separately as biology, chemistry, and physics, whereas it would be taught as one subject in some other federal states. The years of elementary schooling may differ between four and six. In addition, educational objectives and curricula vary considerably (Schecker & Parchmann, 2007).

Tracking. The German education system prior to PISA consisted of three (in some federal states, four) major school tracks. After attending a common elementary school, a student moved on to one of the three school tracks depending on their particular aims and achievement: *Hauptschule*, *Realschule*, or *Gymnasium*. The Hauptschule provided a basic level of education for students up to grade 9 as a preparation for vocational training. The Realschule provided an extended level of education up to grade 10, which allowed students to proceed to higher nonacademic careers, for example, in a small business or bank. Only the Gymnasium provided a high level of education up to grade 12 or 13, which qualified students for attending university. In particular, the education at the Gymnasium centered on the idea of humanistic Bildung. The overarching idea of the three-track system was to provide adequate support according to the different needs of students. In some federal states, an additional fourth track integrated all levels of education in one school form. This track was originally thought to replace the other tracks and, therefore, disestablish the tracking system.

The German Education System in International Comparison

About 5 years before the first PISA study, an empirical comparison of student science achievement in more than 40 countries was provided by the 1995 TIMSS study (Beaton, Martin, Mullis, Gonzalez, Smith & Kelly, 1996). In more than 15,000 schools across the world, students' science achievement was investigated at five grade levels: the two grades with the largest proportion of 9-year-olds (grades 3 and 4), the two grades with the largest proportion of 13-year-olds (grades 7 and 8), and the final grade of secondary education. Although testing of students in grades 7 and 8 was mandatory, the testing of students in other grades was optional. Students' science achievement was evaluated by a written test covering five content dimensions based on an intersection of the participating countries' curricula (earth science, life science, physics, chemistry, and environmental issues) and the nature of science. The results revealed that German students' achievement was average within the participating countries; even worse was that the achievement turned out to be quite heterogeneous among the students (Beaton et al.). Baumert et al. (1997) reported that 20% of the students did not have the science knowledge expected from elementary students and only 25% had an incipient understanding of scientific concepts and processes. Furthermore, the worst result was that the progress of German students from grades 7 to 8 was marginal when compared to their counterparts in other countries (Baumert et al.).

These results were totally unexpected. Educational researchers and policy makers were shaken up; consequently, measures were undertaken to increase empirical research on mathematics and science instruction (e.g. Prenzel et al., 2007) as well as to foster research-based programs to improve school education (e.g. Ostermeier, Prenzel & Duit, 2009). Additionally, policy makers followed the educational researchers' advice and decided to take part in large-scale assessments on a regular basis. Thus, when the idea of a program to compare students' achievement was discussed among the members of the Organisation for Economic Co-operation and Development (OECD), Germany was one of the supporters. The members finally agreed to create the PISA with the goal to assess students' ability to complete tasks relating to real life depending on broad understanding of key concepts of the related science subjects.

German Science Education According to PISA

One of PISA's features is the framework of scientific literacy, which serves as an underpinning of the project (Sadler & Zeidler, 2009). Whereas the TIMSS study focused on how well students master school curriculum, the PISA study "is concerned with the capacity of students to apply knowledge and skills in key subject areas and to analyze, reason and communicate effectively as they pose, solve and interpret problems in a variety of situations" (OECD, 2001, p. 20). This notion of *scientific literacy* aligns with what is termed Vision II of scientific literacy (Roberts, 2007). PISA's definition of scientific literacy and assessments embrace scientific contexts, scientific competencies related to students identifying scientific issues, explaining phenomena scientifically, and using scientific evidence, domains of scientific knowledge, and student attitudes towards science (Bybee, McCrae & Laurie, 2009). However, the PISA notion of scientific literacy does not consider either the development of personality, metalearning, the social or ethical skills central to Bildung (Fischer, Kauertz & Neumann, 2008), or the notion of scientific literacy as a socioscientific issue (Sadler & Zeidler, 2009).

A combined scientific literacy scale was constructed for PISA with a mean score of 500 and a standard deviation of 100 points. Toward the lower end of the scale, students are able to recall simple factual scientific knowledge or use common scientific knowledge in drawing or evaluating conclusions. Toward the upper end of the scale, students are able to create or use conceptual models, make predictions, or give explanations to analyze scientific investigations (OECD, 2001).

Students' science achievement results were reported on this combined scale as well as based on six proficiency levels representing a comprehensive range of scientific literacy (Bybee et al., 2009). Students on the lowest level of proficiency are expected to have only a limited knowledge and be able to apply this knowledge to a limited number of situations. Students on the highest level of proficiency are expected to be consistently able to apply their knowledge to identify scientific issues, explain scientific phenomena, and use scientific evidence. Descriptions of what students can do at each level are based on the content of the items belonging to the respective level (Bybee et al.).

Science Achievement

The PISA 2000 study results confirmed the findings of the TIMSS study: German students' science achievement was significantly below the OECD average and their reading and mathematics performance was just as poor (OECD, 2001). Only a few students (3.4%) mastered the highest proficiency level, while more than one quarter (26.3%) of all students remained on the lowest proficiency level or even below (Prenzel, Rost, Senkbeil, Häußler & Klopp, 2001). The German school system—although making heavy use of tracking—neither reduces inequalities between students nor properly supports them according to their needs. German students' science achievement for 2000 to 2003 improved, in comparison to the other OECD countries, from below the OECD average to the OECD average (OECD, 2004). This result stemmed from a larger number of students mastering the upper proficiency levels, while the number of students scoring on the lowest level of proficiency or even below remained about the same (23.6%); that is, mostly average and above-average students improved (Rost, Walter, Carstensen, Senkbeil & Prenzel, 2004) while the gap between low-proficient and high-proficient students became even wider.

The general trend of improvement continued from 2003 to 2006; German students' science achievement in 2006 was significantly above the OCED average. However, the gap between Germany and the top-scoring countries was still considerably large (between 15 and 47 points on the PISA scientific literacy scale). Slightly < 20% of all students were on the lowest proficiency level, while 1.8% of all students were on the highest proficiency level. Taking into account that six proficiency levels were defined and that 11.8% of the German students were found on the two highest levels (Prenzel et al., 2007), it seemed that students of all proficiency levels improved. A subset of items was the same from 2000 to

2006, which allowed a comparison of students' science achievement across the test administrations. The improvement in achievement can be considered substantial although the variance had not been considerably reduced with the passage of time (Prenzel, Rönnebeck, & Carstensen 2008; see also Carstensen, Prenzel & Baumert, 2008).

Science Achievement in the Federal States

Within the scope of the PISA study, a national add-on study was conducted to compare students' science achievement across the 16 federal states. The national study built upon the same instruments and procedures of analysis that had been used in the international study. Whereas the German sample for the international comparisons embraced about 4,500 students, the sample for the national comparisons embraced about ten times as many students, which was necessary to be able to compare the achievements across the federal states. The findings from the 2000 national add-on study revealed particular differences in student achievement. State average performances ranged from well below to slightly above the PISA average (Prenzel, Carstensen, Rost & Senkbeil, 2002). The national add-on studies in 2003 and 2006 confirmed the findings from the international study. Student achievement improved from 2000 to 2003 in all federal states, but again the growth was found to stem almost exclusively from average and above-average performing students. The percentage of scientifically illiterate students did not improve. Moreover, while the trailing states scored below the OECD average, the leading states scored well above (Rost, Senkbeil, Walter, Carstensen & Prenzel, 2005). The national add-on study in 2006 showed that the best states had closed the gap to the top countries in the international ranking. However, most states still scored in the OCED average and one state performed significantly below the average (Rönnebeck, Schöps, Prenzel & Hammann, 2008).

Science Achievement and System Characteristics

Since extensive additional data (e.g. students' social background, different perceptions of and attitudes toward teaching and learning) were collected within PISA, it is possible to link students' science achievement to student, home, and school traits (Anderson, Lin, Treagust, Ross & Yore, 2007). A central finding for Germany was the substantial influence of the socioeconomic status (SES) on student achievement (OECD, 2001). Additionally, social background differed considerably between school tracks; students at Hauptschule were mostly from working class families, whereas most of the students at Gymnasium were from middle and upper-middle class families or families with an even higher social status (Baumert & Schümer, 2001). The slope of the social gradient and the magnitude of the relation between SES and science achievement were particularly great although close to the average of all OECD countries (Ehmke & Baumert, 2007). As a tracking school system is supposed to foster students according to their needs and independent from their social background, it should reduce social disparities and lead to a less-than-average relation between parents' SES and students' science achievement. It seems that the three-track school system has failed in that objective.

Science Education Reform in Germany

The international large-scale assessments in TIMSS and PISA revealed the reality of science education in Germany: students had an average performance in international comparisons and the particular objectives of science education central to these assessments were not achieved. At the beginning of a millennium that was characterized by growing relevance of science and technology for everyday life, one quarter of the German students failed to achieve a reasonable level of scientific literacy. Moreover, disparity between students of different socioeconomic backgrounds was not reduced by the three-track system. Instead, regional and social differences were found for student science achievement—the best students scored in the range of the top countries' scores, while the poor students' achievement accounted for the less-than-expected average overall achievement. These results were taken as evidence that minimum standards were needed in Germany (Klieme et al., 2003).

It became obvious from the PISA studies that to close the gap between Germany and the top countries would require more than occasional attempts at reform and quality improvement and that more comprehensive measures were needed. Therefore, policy makers decided for a fundamental reform of the educational system from one in which educational policy is implemented by exerting control over the input to one in which educational policy is implemented by defining the output. The core piece of this reform was the introduction of the NES for the end of lower secondary education (grade 10, students about 16 years old) as a basis to develop an assessment system that allows monitoring the quality of the educational system continuously along further reform (Klieme et al., 2003).

Naturwissenschaftliche Grundbildung

The introduction of the NES for science education was accompanied by an extensive discussion on the state and perspective of German science education (Klieme et al., 2003). In particular, the advantages of the notion of humanistic Bildung over the economically oriented notion of scientific literacy underpinning the PISA studies were highly debated. It was feared that the exclusive reliance on PISA results for policy decisions would lead to a standardization of education that in turn would mean the sacrifice of Bildung in favor of a pragmatic concept of testable, utilitarian knowledge (Jürgens, 2004).

On the other hand, it became clear that internationally scientific literacy, albeit in many notions, had become a synonym for the purpose of science education. Klieme et al. (2003) identified three positions in the discussion on the objectives of education: (a) the position of Bildung favoring the specific German notion of education, (b) the literacy position favoring the Anglo-American notion of education, and (c) the position of fundamental abilities that accepts the expectations society has toward the education system and the necessity of standards as an operational definition of these expectations. The *Naturwissenschaftliche Grundbildung* emerged, which incorporates the notion of scientific literacy as used in PISA. Naturwissenschaftliche Grundbildung emphasizes "science in life situations in which science plays a key role" (Bybee et al., 2009, p. 866) and in science-related situations that students are likely to encounter as citizens (Roberts, 2007). Moreover, it embraces understandings and practices pertinent to science-related situations in which individuals integrate science with individual and social considerations similar to the socioscientific notion of scientific literacy (Sadler & Zeidler, 2009). Therefore, Naturwissenschaftliche Grundbildung became the basis for the formulation of the NES in the science subjects.

National Education Standards

The introduction of the NES for grade 10, the end of lower secondary level in science education, served as a vision and a measure that would bring an end to the diversity of curricula across the 16 federal states, decrease social injustice, and make classroom teaching more effective (Nentwig & Schanze, 2007). Since science was mostly taught as the three individual subjects of biology, chemistry, and physics, NES were developed for each subject. By order of the Ständige Konferenz der Kultusminister (i.e. the ministers of education and cultural affairs of each federal state) and under participation of all federal states, standards for primary school qualification in German and mathematics and for middle school qualification in German, mathematics, English, and a little later, biology, chemistry, and physics were published in 2004. These standards marked the levels of ability that students should reach in a subject at the end of grade 4 and after 10 years of participation in the school system (Klieme et al., 2003).

The NES are addressed to teachers, learners, and institutions. This approach represents a dramatic change from past practice; it demands that schools and the school system provide students with opportunities to learn to meet the standards defined. In turn, national assessments based on the standards are expected to provide information where the school system and schools fail to foster students' learning so that they can meet the standards. This is thought to be a central element of quality assessment and subsequent quality development.

The standards have the same form for all three subjects: a preamble outlines the particular contribution of the respective subject to students' Naturwissenschaftliche Grundbildung. Like the PISA framework, the standards differentiate the components of *Naturwissenschaftliche Grundbildung*: content knowledge (phenomena, concepts, principles, facts, laws, and basic concepts) and scientific inquiry (knowledge about scientific inquiry, the nature of science, and experimentation skills). Naturwissenschaftliche Grundbildung extends the notion of scientific literacy used in PISA by adding two components: communication (communicating information in accordance with the content and audience) and assessment (using adequate criteria for assessment based on scientific facts and cases). The content knowledge defines unified themes or basic concepts (Bybee, 1998). Liu (2009) pointed out that organizing standards "by unified themes means that the same theme or concept may be learned again and again at different grades with increasing complexity and thus increasingly higher expectations of proficiency" (p. 8). The basic concepts in biology are system, structure and function, and evolution; in chemistry, matter–particle relations, structure–property relations, chemical reactions, and energy; and in physics, matter, interaction, system, and energy.

For each component of Naturwissenschaftliche Grundbildung defined, a set of competencies is listed in the NES (Table 1). Competencies are understood as the cognitive and affective abilities as well as the motivational, volitional, and social readiness to solve problems in variable situations (Weinert, 2001). The German NES may be considered to be performance standards unlike the standards in the US, which focus on key content by "specifying what facts, concepts and forms of inquiry should be learned and how they should be taught and evaluated" (Eisenhart, Finkel & Marion, 1996, p. 266). Furthermore, the US standards mostly support scientific literacy Vision I (Roberts, 2007), while each individual component of Naturwissenschaftliche Grundbildung appears to be related to scientific literacy Vision II.

The conception of the NES as performance standards required that proficiency levels be established as the basis on which student performance could actually be measured. This is the same conception as in PISA where proficiency levels were described with respect to the different components of scientific literacy. The description of these levels in PISA were obtained by grouping items within particular ranges of difficulty and describing the levels based on the content of these items with respect to the framework of scientific literacy. However, during PISA 2000, attempts to validate these proficiency levels failed (Prenzel et al., 2001); and, at the time of the formulation of the NES for the end of lower secondary level education in the science subjects, no validated hierarchy of proficiency levels existed. Therefore, different requirements were specified, which were explicitly characterized as nonhierarchical. The requirements were termed reproduction, application, and transfer and were based on Bloom's taxonomy (Anderson & Krathwohl, 2001). In summary, the German NES describe four components and requirements with respect to these components that result in a two-dimensional model of Naturwissenschaftliche Grundbildung, which could be referred to as a model of competency.

TABLE 1

Example competencies from the German NES for physics

Area of competency	*Students will …*
Content knowledge	Possess a profound, well-structured knowledge based on the basic concepts.
	Use analogies to solve tasks and problems.
Scientific inquiry	Formulate hypotheses at hand of ordinary examples.
	Plan, perform, and analyze ordinary experiments.
Communication	Research a topic using different sources.
	Present results of their work in a way appropriate to the audience.
Assessment	Compare and judge alternative technical solutions considering physical, economic, social, and ecological aspects.
	Name implications of physics-related findings with respect to historical and social contexts.

Models of Competency

A competency model describes areas of competency and respective proficiency levels, so-called competency levels; builds the missing link between the abstract formulation of performance standards and benchmarks of the standards; and allows for identifying which level students have mastered for target competencies. Consequently, based on a competency model, information can be obtained to identify which educational objectives have been achieved and which obstacles need to be overcome to achieve other objectives—for an individual student, a classroom of students, or a whole population of students in a state or a country. Models of competency, therefore, help in the difficult task of measuring standards. Without doubt, there is a particular amount of expertise in the field of assessing students' knowledge; but the question of how to measure students' mastery of standards, in particular those building on the notion of scientific literacy Vision II, still remains largely unanswered (Fensham, 2009a). Teachers in schools as well as science education reformers need more detailed information about what students individually have already mastered and what their next level of development could be. That is, qualitatively different levels of competency that build on each other are required.

Not including a hierarchy of competency levels in the NES was a design flaw right from the beginning. Consequently, the deliberations of this particular issue since the publication of the NES have resulted in a variety of different models being suggested and empirically investigated for biology (Mayer, Grube & Möller, 2008), chemistry (Bernholt, Parchmann & Commons, 2009), and physics (Kauertz & Fischer, 2006; Schecker & Parchmann, 2006). Several researchers were successful in providing evidence that a hierarchy of competency levels can be described based on the complexity of the particular knowledge that needs to be applied in a given situation (Bernholt et al.; Kauertz & Fischer). Kauertz & Fischer provided evidence that a system of six levels of complexity between factual knowledge and understanding of scientific concepts may be used to describe students' competency with respect to related content knowledge. This model was the first that could be empirically validated as a hierarchy of competency levels, and it served as a point of departure for the specification of a competency model used to benchmark and subsequently adjust the NES (Fig. 1). This model was extended to describe levels of competency for content knowledge and inquiry for biology, chemistry, and physics (Walpuski, Kampa, Kauertz & Wellnitz, 2008). A nationwide validation of the model has just been completed. (As data collection has just been completed at the end of 2009, no results have been published yet. Further information can be obtained from http://www.iqb.hu-berlin.de/arbbereiche/projekte?pg=p_34, in German.)

Furthermore, the model will be extended to the remaining two areas of competency (communication and assessment) and validated. Benchmarking of the NES with a representative sample of German students is planned for 2011. This will provide information that will allow for adjusting the NES.

Summary and Outlook

Before the PISA study, the German education system was characterized by its own particular notion of education: humanistic Bildung. National standards and national assessments were not issues in educational policy. However, when PISA findings revealed particular deficits in German science education, an extensive and controversial discussion emerged on the state and perspective of science education. Policy makers decided for a structural reform of the educational system, at the heart of which was the formulation of the NES in the form of performance standards. Based on the notion of Naturwissenschaftliche Grundbildung, which in essence is an extended notion of scientific literacy underpinning the PISA framework, the NES for science education were developed. The standards

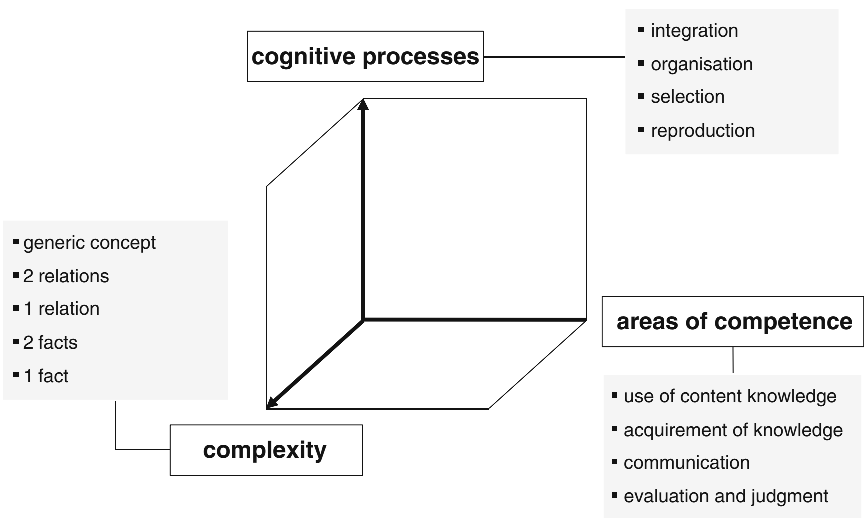

Figure 1. Competency model used to benchmark and adjust the German NES for science education

were framed by the PISA framework and the particular deficits of German science education identified by PISA. Not surprisingly, the NES for science education have a similar structure as PISA; competencies are described for four components of Naturwissenschaftliche Grundbildung, and performance requirements are described for the competencies. The lack of proficiency levels in the standards provided an opportunity and promoted a completely new field of research: competency modeling. The findings of this research provided a foundation for models of competency specified for the three science subjects. Based on this model, the NES for the science subjects will be benchmarked in 2011, following which it will be possible to adjust the normatively set standards on empirical evidence.

The time frame (1995–2011+) reveals that developing standards and measuring student achievement with respect to these standards is a challenging task. The development of standards is an iterative process in which standards are, first, described normatively based on educational theories and existing empirical knowledge; second, benchmarked with results from large, representative samples; and third, adjusted to reflect this evidence—just to be benchmarked again. The formulation of standards is a normative process that is highly influenced by the educational objectives defined within society; standards might even change after they are empirically fine tuned. Examining the benchmarking results will provide information on possible areas to strengthen science education because science education "ought to provide opportunities for learners to experience science in contexts analogous to the contexts that they may confront in their lived experiences" (Sadler & Zeidler, 2009, p. 912); that is, contexts in which they are required to apply the acquired competencies and which might not be known at the time of schooling. The process of developing and benchmarking standards in the way it happened in Germany is a prototypical case of policy making informed by research. Fensham (2009b) argued, "if research has any value, it should be part of the process of making policy in education" (p. 1091).

However, there are still several challenges for Germany to meet in the process of standards development and benchmarking. First, the overall consequence of this change in policy making needs to be evaluated. That is, does the reform of science education, more specifically the introduction of the NES, actually improve student achievement in science? Will further measures in the scope of the ongoing reform of the education system improve education in general and in science education in particular? There is room for improvement of the test instruments used for benchmarking. It is well understood how to assess students' knowledge but little is known about how to assess competencies. How can we ensure that tasks do not assess only students' knowledge, which certainly is a requirement for competency? What particular influence is the interaction between everyday context and science content? How can tasks and contexts be designed so that they create a situation that is as close as possible to real-life situations and still be related to Naturwissenschaftliche Grundbildung?

References

Anderson, J. O., Lin, H.-L., Treagust, D. F., Ross, S. P., & Yore, L. D. (2007). Using large-scale assessment datasets for research in science and mathematics education: Programme for International Student Assessment (PISA). *International Journal of Science and Mathematics Education, 5*(4), 591–614.

Anderson, L. W., & Krathwohl, D. R. (2001). *A taxonomy for learning, teaching, and assessing: A revision of Bloom's taxonomy of educational objectives.* New York: Longman.

Baumert, J., Lehmann, R., Lehrke, M., Schmitz, B., Clausen, M., Hosenfeld, I., et al. (1997). *TIMSS—Mathematisch-Naturwissenschaftlicher Unterricht im internationalen Vergleich* [TIMSS—Mathematic and scientific instruction in international comparison]. Opladen, Germany: Leske + Buderich.

Baumert, J., & Schümer, G. (2001). Familiäre Lebensverhältnisse, Bildungsbeiteiligung und Kompetenzerwerb [Family life situation, educational participation, and competency acquisition]. In J. Baumert, E. Klieme, M. Neubrand, M. Prenzel, U. Schiefele, W. Schneider, et al. (Eds.), *PISA 2000. Basiskompetenzen von Schülerinnen und Schülern im internationalen Vergleich* [PISA 2000: Basic compencies of students in international comparison] (pp. 324–410). Opladen, Germany: Leske + Budrich.

Beaton, A. E., Martin, M. O., Mullis, I. V., Gonzalez, E. J., Smith, T. A., & Kelly, D. L. (1996). *Science achievement in the middle school years: IEA's Third International Mathematics and Science Study (TIMSS).* Chestnut Hill, MA: TIMSS International Study Center, Boston College.

Bernholt, S., Parchmann, I., & Commons, M. L. (2009). Kompetenzmodellierung zwischen Forschung und Unterrichtspraxis [Modelling competencies between research and instructional practice]. *Zeitschrift für Didaktik der Naturwissenschaften, 15*, 219–245.

Bybee, R. W. (1998). National standards, deliberation, and design: The dynamics of developing meaning in science curriculum. In D. A. Roberts & L. Ostman (Eds.), *Problems of meaning in science curriculum* (pp. 150–165). New York: Teachers College Press.

Bybee, R. W., McCrae, B., & Laurie, R. (2009). PISA 2006: An assessment of scientific literacy. *Journal of Research in Science Teaching, 46*(8), 865–883.

Carstensen, C. H., Prenzel, M., & Baumert, J. (2008). Trendanalysen: Wie haben sich die Kompetenzen in Deutschland zwischen PISA 2000 und PISA 2006 entwickelt? [Trend analyses: How have competencies in Germany developed from PISA 2000 to PISA 2006? Special Issue]. *Zeitschrift für Erziehungswissenschaft*, *10*, 11–34.

Döbert, H. (2007). Germany. In W. Hörner, H. Döbert, B. von Knopp, & W. Mitter (Eds.), *The education systems of Europe* (pp. 299–325). Amsterdam, The Netherlands: Springer.

Ehmke, T., & Baumert, J. (2007). Soziale Herkunft und Kompetenzerwerb: Vergleiche zwischen PISA 2000, 2003 und 2006 [Social background and competence acquisition: Comparisons between PISA 2000, 2003 and 2006]. In M. Prenzel, C. Artelt, J. Baumert, W. Blum, M. Hammann, E. Klieme, et al. (Eds.). *PISA 2006: Die Ergebnisse der dritten internationalen Vergleichsstudie* [PISA 2006: Results of the third international comparison study] (pp. 309–336). Münster, Germany: Waxmann.

Eisenhart, M., Finkel, E., & Marion, S. F. (1996). Creating the conditions for scientific literacy: A re-examination. *American Educational Research Journal, 33*, 261–295.

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

Fensham, P. J. (2009b). The link between policy and practice in science education: The role of research. *Science Education, 93*(5), 1076–1095.

Fischer, H. E., Kauertz, A., & Neumann, K. (2008). Standards of science education. In S. Mikelskis-Seifert, U. Ringelband, & M. Brückmann (Eds.), *Four decades of research in science education—From curriculum development to quality improvement* (pp. 29–41). Münster, Germany: Waxmann.

Jürgens, E. (2004). Pädagogische Implikationen der KMK-Entwürfe für Bildungsstandards [Pedagogical implications of the KMK drafts of education standards]. In J. Schlömerkemper (Ed.), *Bildung und Standards* (pp. 48–65). Weinheim, Germany: Juvena.

Kauertz, A., & Fischer, H. E. (2006). Assessing students' level of knowledge and analysing the reasons for learning difficulties in physics by Rasch analysis. In X. Liu & W. Boone (Eds.), *Applications of Rasch measurement in science education* (pp. 212–246). Maple Grove, MA: Jam Press.

Klafki, W. (1996). *Neue Studien zur Bildungstheorie und Didaktik* [New studies on educational theory and didactics] (5th ed.). Weinheim, Germany: Beltz.

Klieme, E., Avenarius, H., Blum, W., Döbrich, P., Gruber, H., Prenzel, M., et al. (2003). *Zur Entwicklung nationaler Bildungsstandards* [Regarding the development of National Education Standards]. Berlin, Germany: Bundesministerium für Bildung und Forschung.

Liu, X. (2009). *Linking competencies to opportunities to learn.* New York: Springer.

Mayer, J., Grube, C., & Möller, A. (2008). Kompetenzmodell naturwissenschaftlicher Erkenntnisgewinnung [A competency model of scientific inquiry]. In U. Harms & A. Sandmann (Eds.), *Lehr- und Lernforschung in der Biologiedidaktik* (vol. 3, pp. 63–80). Innsbruck, Austria: Studienverlag.

Nentwig, P., & Schanze, S. (2007). Making it comparable—Standards in science education. In D. Waddington, P. Nentwig, & S. Schanze (Eds.), *Standards in science education* (pp. 11–19). Münster, Germany: Waxmann.

Organisation for Economic Co-operation and Development (2001). *Knowledge and skills for life: First results from PISA 2000.* Paris: Author.

Organisation for Economic Co-operation and Development (2004). *Learning for tomorrow's world: First results from PISA 2003*. Paris: Author.

Ostermeier, C., Prenzel, M., & Duit, R. (2009). Improving science and mathematics instruction: The SINUS Project as an example for reform as teacher professional development. *International Journal of Science Education, 32*(3), 303–327.

Prenzel, M., Carstensen, C. H., Rost, J., & Senkbeil, M. (2002). Naturwissenschaftfliche Grundbildung im Ländervergleich [Scientific literacy compared across the federal states]. In J. Baumert, C. Artelt, E. Klieme, M. Neubrand, M. Prenzel, U. Schiefele, et al. (Eds.). *PISA 2000—Die Länder der Bundesrepublik Deutschland im Vergleich* [PISA 2000—The federal states of the Federal Republic of Germany compared] (pp. 129–158). Opladen, Germany: Leske + Budrich.

Prenzel, M., Rönnebeck, S., & Carstensen, C. H. (2008). PISA 2006—Eine Einführung in den Ländervergleich [PISA 2006—An introduction to the comparison of the federal states]. In M. Prenzel, C. Artelt, J. Baumert, W. Blum, M. Hammann, E. Klieme, et al. (Eds.). *PISA 2006 in Deutschland. Die Kompetenzen der Jugendlichen im dritten Ländervergleich* [PISA 2006 in Germany: Teenagers' competencies in the third comparison of the federal states] (pp. 31–64). Münster, Germany: Waxmann

Prenzel, M., Rost, J., Senkbeil, M., Häußler, P., & Klopp, A. (2001). Naturwissenschaftliche Grundbildung: Testkonzeption und Ergebnisse [Scientific literacy: Test conception and results]. In J. Baumert, E. Klieme, M. Neubrand, M. Prenzel, U. Schiefele, W. Schneider, et al. (Eds.) *PISA 20000. Basiskompetenzen von Schülerinnen und Schülern im internationalen Vergleich* [PISA 2000: Basic compencies of students in international comparison] (pp. 191–248). Opladen, Germany: Leske + Budrich.

Prenzel, M., Schöps, K., Rönnebeck, S., Senkbeil, M., Walter, O., Carstensen C. H., & Hammann, M. (2007). Naturwissenschaftliche Kompetenz im internationalen Vergleich [Scientific competency in international comparison]. In M. Prenzel, C. Artelt, J. Baumert, W. Blum, M. Hammann, E. Klieme, et al. (Eds.). *PISA 2006: Die Ergebnisse der dritten internationalen Vergleichsstudie* [PISA 2006: Results of the third international comparison study] (pp. 63–106). Münster, Germany: Waxmann.

Roberts, D. A. (2007). Scientific literacy/science literacy. In S. K. Abell & N. G. Lederman (Eds.), *Handbook of research on science education* (pp. 729–780). Mahwah, NJ: Lawrence Erlbaum.

Rönnebeck, S., Schöps, K., Prenzel, M., & Hammann, M. (2008). Naturwissenschaftliche Kompetenz im Ländervergleich [Scientific competency in comparison of the federal states]. In M. Prenzel, C. Artelt, J. Baumert, W. Blum, M. Hammann, E. Klieme, et al. (Eds.). *PISA 2006 in Deutschland. Die Kompetenzen der Jugendlichen im dritten Ländervergleich* [PISA 2006 in Germany: Teenagers' competencies in the third comparison of the federal states] (pp. 67–94). Münster, Germany: Waxmann.

Rost, J., Senkbeil, M., Walter, O., Carstensen, C. H., & Prenzel, M. (2005). Naturwissenschaftliche Grundbildung im Ländervergleich [Scientific literacy compared across the federal states]. In M. Prenzel, J. Baumert, W. Blum, R. Lehmann, D. Leutner, M. Neubrand, et al. (Eds.). *PISA 2003. Der zweite Vergleich der Länder in Deutschland* [PISA 2003: The second comparison of the German federal states] (pp. 103–124). Münster, Germany: Waxmann.

Rost, J., Walter, O., Carstensen, C.H., Senkbeil, M., & Prenzel, M. (2004). Naturwissenschaftliche Kompetenz [Scientific competency]. In M. Prenzel, J. Baumert, W. Blum, R. Lehmann, R. D. Leutner, M. Neubrand, et al. (Eds.). *PISA 2003. Der Bildungsstand der Jugendlichen in Deutschland—Ergebnisse des zweiten internationalen Vergleichs* [PISA 2003: Results of the second international comparison] (pp. 111–146). Münster, Germany: Waxmann.

Sadler, T. D., & Zeidler, D. L. (2009). Scientific literacy, PISA and socioscientific discourse: Assessment for progressive aims of science education. *Journal of Research in Science Teaching, 46*(8), 909–921.

Schecker, H., & Parchmann, I. (2006). Modellierung naturwissenschaftlicher Kompetenz [Modelling scientific literacy]. *Zeitschrift für die Didaktik der Naturwissenschaften, 12*, 45–66.

Schecker, H., & Parchmann, I. (2007). Standards and competence models: The German situation. In D. Waddington, P. Nentwig, & S. Schanze (Eds.), *Standards in science education* (pp. 147–164). Münster, Germany: Waxmann.

Sorkin, D. (1983). Wilhelm von Humboldt: The theory and practice of self-formation (Bildung). *Journal of the History of Ideas, 44*, 55–73.

Ständige Konferenz der Kultusminister der Länder der Bundesrepublik Deutschland [Standing Conference of the Ministers of Education and Cultural Affairs of the Länder in the Federal Republic of Germany] (1997). Pressemitteilungen vom 27.06. und 24.10.1997 zu den Plenarsitzungen Nr. 279/280 der Kultusministerkonferenz am 26./27.06 und 23./24.10.1997 [Press release from 27.06 and 24.10.1997 concerning the plenary meetings no. 279/280 of the Conference of the Ministers of Education and Cultural Affairs on 26./27 and 23./24.10.1997]. Bonn, Germany: Author.

Walpuski, M., Kamper, N., Kauertz, A., & Wellnitz, N. (2008). Evaluation der Bildungsstandards in den Naturwissenschaften [Evaluation of the National Education Standards in the sciences]. *Mathematisch und Naturwissenschaftlicher Unterricht, 6*, 223–226.

Weinert,, W 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). Seattle, WA: Hogrefe & Huber.

Knut Neumann
Department of Physics Education
Leibniz-Institute for Science and Mathematics Education (IPN)
Kiel, Germany
E-mail: neumann@ipn.uniiel.de

Hans E. Fischer
University Duisburg–Essen
Essen, Germany

Alexander Kauertz
University of Education of Weingarten
Weingarten, Germany