Chapter 10
Large-Scale Studies in Mathematics Education Research

Kristina Reiss, Andreas Obersteiner, Aiso Heinze, Ursula Itzlinger-Bruneforth and Fou-Lai Lin

Abstract Large-scale studies assess mathematical competence in large samples. They often compare mathematical competence between groups of individuals within or between countries. Although large-scale research is part of empirical educational research more generally, it is also linked to more genuine mathematics education research traditions, because sophisticated methods allow for empirical verifications of theoretical models of mathematical competence, and because results from large-scale assessments have influenced mathematics education practices. This chapter provides an overview of large-scale research in mathematics education in German speaking countries over the last decades. After a brief review of historical developments of large-scale assessments in Germany, we focus on the development of competence models in Germany and Austria. At the end of this chapter, we reflect on recent developments and discuss issues of large-scale assessments more generally, including an international perspective.

Keywords Large-scale assessment · Mathematical competence · Competence models · PISA study

K. Reiss (✉)
TUM School of Education, Technical University of Munich, Munich, Germany
e-mail: kristina.reiss@tum.de

A. Obersteiner
Freiburg University of Education, Freiburg im Breisgau, Germany

A. Heinze
Leibniz Institute for Science and Mathematics Education, Kiel, Germany

U. Itzlinger-Bruneforth
Federal Institute for Educational Research, Innovation & Development of the Austrian School System, Salzburg, Austria

F.-L. Lin
National Taiwan Normal University, Taipei, Taiwan

© The Author(s) 2019

H. N. Jahnke and L. Hefendehl-Hebeker (eds.), *Traditions in German-Speaking Mathematics Education Research*, ICME-13 Monographs,
https://doi.org/10.1007/978-3-030-11069-7_10

10.1 Introduction

In recent years, researchers have assessed mathematical competence in large samples using sophisticated statistical methods. Large-scale assessment of mathematical competences requires close collaboration between researchers from mathematics education, statistics, and psychology. Accordingly, the lines of research presented in this chapter are linked to both mathematics education research traditions and, more broadly, large-scale empirical educational research. The chapter is structured into five sections. The following paragraphs provide an introduction to the contents of the five sections.

Popular large-scale studies such as PISA (Programme for International Student Assessment) have received much public attention and have led to a stronger focus on the outcome of school education in general and of mathematics education in particular. In Germany, international large-scale studies did not receive much attention before 1995, when Germany took part in the Third International Mathematics and Science Studies (TIMSS) for the first time. The results showed that German lower and upper secondary school students' mathematical performance did not meet the expectations of teachers, educators, and the general public. German students performed below the international average and showed acceptable results only for routine problems (Baumert et al. 1997). The results of PISA 2000 (Baumert et al. 2001) were again disappointing. Consequences from these studies were intensive debates among educators and stakeholders, and the launch of educational programs to improve mathematics instruction at school. Another consequence was the agreement to use large-scale assessments on a regular basis to monitor the outcome of school education. We elaborate on these developments in Sect. 10.2.

The developments in large-scale empirical educational research are related to mathematics education research because assessing mathematical competence in large samples was accompanied by the development and the empirical verification of theoretical models of mathematical competence. Assessing students' mathematical competences requires models of what mathematical competence should be. Initial models were predominantly based on theoretical and normative considerations but rarely on empirical evidence. Reiss and colleagues (e.g., Reiss et al. 2007a, b; Reiss et al. 2012) developed a model for primary mathematics education that took into account theoretical and normative perspectives and was continuously refined on the basis of empirical evidence. The model suggests five levels of mathematical competence ranging from technical background knowledge and routine procedures to complex mathematical modelling. Section 10.3 of this chapter provides an overview of this model of mathematical competence.

To monitor the outcome of education on a regular basis, new institutions were founded in Germany, such as the Institute for Educational Quality Improvement (IQB, Berlin) and the Center for International Student Assessment (ZIB, Munich). The idea of system monitoring is, however, not specific to Germany. Other countries founded similar institutions and developed similar models of mathematical competence to assess students' competences on a regular basis. In Austria, the Federal Institute for

Educational Research, Innovation and Development of the Austrian School System (BIFIE) was founded on a similar basis as institutions in Germany. One core function of the institute was to establish and disseminate knowledge about educational standards in mathematics. These standards were based on a theoretical framework which included a model of mathematical competence with three dimensions (process, content, and complexity). This model is based on the existing mathematics curriculum and represents a new structure for teaching, learning and assessment of mathematics. In Sect. 10.4, we elaborate on this model of mathematical competence and its development in Austria.

In conclusion, large-scale studies allow monitoring the outcome of mathematics education on the system level. The broad data these studies provide have been used to empirically validate theoretical models of mathematical competence and have contributed to a more realistic view on what students are capable of learning at school. However, there are several general issues of large-scale studies such as their purpose and their risks that need to be discussed. At the end of this chapter, in Sect. 10.5, we reflect on recent developments and discuss more general issues of large-scale assessments that reach beyond the traditions in German-Speaking countries.

10.2 Large-Scale Assessment: Impact on Mathematics Educational Research and Practice in Germany

The aim of this section is to provide an overview of the impact of large-scale assessment (LSA) studies on mathematics educational research and practice in Germany. The section is structured into four subsections. The first subsection provides some historical background information about the public discussion of the LSA results from TIMSS 1995 and PISA 2000 in Germany. In the next three subsections, we describe the impact of LSA on mathematics educational research, on mathematics education as a scientific discipline, and on schools and mathematics teachers, respectively.

10.2.1 LSA of Mathematical Competence in Germany: How It Started (TIMSS 1995 and PISA 2000)

The history of (international) LSA of mathematical competence started in Germany in 1995. Until then, Germany did not take part in international comparison studies on outcomes of mathematics education. Moreover, until 2000 there were no comparisons between the 16 federal states in Germany. The latter is particularly interesting because in the 1970s and 1980s there were heated (political) debates within and between the federal states about the most effective school system, the educational goals, and the preferred teaching styles. Although there was disagreement among

the federal states, politicians were convinced that Germany in general had a high-level educational system. Hence, many were disappointed by the results of TIMSS 1995 for lower and upper secondary students in mathematics. Germany scored below the international average and the performance level was only acceptable for routine problems (Baumert et al. 1997). The consequence was a discussion within the educational sciences and educational administrations (Blum and Neubrand 1998) but there was hardly any discussion in the public. The results of TIMSS 1995 were supplemented by the findings of TIMSS Video 1995. This study confirmed findings from earlier case studies (e.g., Bauersfeld 1978) describing a typical teaching style in German mathematics classrooms (Stigler et al. 1999; Neubrand 2002). A feature of this teaching style was a low variation of teaching methods, which often meant direct instruction accompanied by low-level question-answer-sequences (so-called "funnel pattern", Bauersfeld 1978). Moreover, a low quality of tasks and task implementation was reported. As a consequence, the responsible politicians decided in 1997 to participate in international comparisons and to compare educational outcomes of the 16 federal states on a regular basis.

In 2000, Germany took part in the first PISA study. Like the results of TIMSS, the results of PISA were again relatively weak, from both a criterion-oriented and an international comparative perspective (Deutsches PISA-Konsortium 2001). For mathematics, the study identified a large group of low-performing students (about 25%) and only a small group of high-performing students (about 2%). Moreover, there was a strong relation between students' achievement and their socio-economic background, and there were large differences between the 16 German federal states. A specific weakness of German students was the ability to use their mathematical knowledge in real-life situations. In contrast to the TIMSS 1995 results, the PISA results were debated in the broader public. For example, the fact that German students and students from the U.S. were quite similar with respect to the level of mathematics competence and the relation between achievement and socio-economic background was discussed intensively.

10.2.2 Impact on (Mathematics) Educational Research in Germany

The political decision to participate in international LSA provided various opportunities for educational research. Accordingly, the LSA itself as well as the results of the LSA had an impact on (mathematics) educational research in Germany. On the one hand, it influenced descriptive research, and on the other hand, it induced studies aiming at explanations for the descriptive LSA findings.

The motivation for further descriptive research studies relied on the mission of educational research to provide a basis for political decisions to improve the educational system. With TIMSS 1995 and PISA 2000 it became clear that educational policy in Germany was based on insufficient information about the educational real-

ity. Hence, subsequent to the first international LSA on the secondary level, educational research used the opportunities of national extensions of the follow-up cycles as well as of additional LSA in other age groups. For example, in PISA 2003, the German sample was supplemented by an additional sample of students with migration background (so-called oversampling). It turned out that language skills and socio-economic status (SES) were the main factors for the comparatively low mathematics performance of students with migration background. Moreover, the PISA 2003 sample was tested again one year later to collect information about students' learning progress in mathematics. The findings indicated that there was large variation in mathematics competence development between school classes. Only 58% of students showed a significant increase of mathematical competence from grade 9 to grade 10 on the mathematics PISA scale (for more details on how mathematical competence was conceptualized, see Sect. 10.3.1).

Several additional LSA studies were conducted with respect to mathematical competence in other populations than secondary students, such as:

- TIMSS in grade 4 at the end of German primary school. The findings for mathematics competence were similar to those observed in secondary school. This means that many problems already evolve in primary schools.
- TEDS (Teacher Education and Development Study)—the study on professional knowledge of mathematics pre-service teachers (Blömeke et al. 2010). One of the striking results was that non-certified mathematics teachers show very low professional knowledge.
- PIAAC (Programme for the International Assessment of Adult Competencies)—the study of adults' competence in various domains including mathematics. The mathematical competence of German adults is on an average level in comparison to the OECD countries. However, about 17% of the German adults show very poor mathematical competence, which corresponds to the PISA results (Rammstedt 2013). Typically, adults master mathematical procedures they need in their every-day or professional live (the "use it or loose it phenomenon", Duchhardt et al. 2017).
- NEPS (National Educational Panel Study). This project started in 2008 with the goal of creating a panel of longitudinal data of competence development (mathematics, science, language) and specific conditions of learning environments (www.neps-data.de). The data collection is based on a multi-cohort-sequence-design which follows cohorts starting with babies, kindergarten children, 5th-graders, 9th-graders, university students, and adults. The panel provides data for the scientific community which allows examining longitudinal competence development in mathematics (Neumann et al. 2013).

In addition to descriptive research several explanatory research studies and study programs were conducted. In essence, the aim of these projects was to explain the previously obtained descriptive results and to generate knowledge which might help to improve mathematics education in schools. Examples for research programs are the priority programs "Educational Quality of Schools" and "Competence Models", supported by the German Science Foundation (DFG), or the framework program

"Empirical Educational Research", supported by the Federal Ministry of Education and Research. Examples of studies that were conducted within these programs are:

- The Pythagoras video study (Lipowsky et al. 2009) focusing on lessons about the Pythagorean Theorem. This study combined data on the quality of mathematics instruction with data on students' learning progress.
- Projects on reasoning and proof in geometry classroom (Reiss et al. 2007a) or teaching problem solving (Komorek et al. 2007). These projects analyzed conditions for successful mathematics instruction fostering students' competences in geometric proof and in problem solving. They also developed teaching concepts and material and evaluated them in intervention studies.
- The COACTIV study (Professional Competence of Teachers, Cognitively Activating Instruction, and Development of Students' Mathematical Literacy), which investigated professional competence of mathematics teachers and aspects of their mathematics instruction in the PISA 2003 follow-up sample (Kunter et al. 2013). This study provided evidence for the influence of mathematics teachers' content knowledge (CK) and pedagogical content knowledge (PCK) on instructional quality and students' learning progress. In particular, the complexity of mathematical problems used in German mathematics classrooms within one school year was examined (Jordan et al. 2008). The researchers evaluated 47,500 problems from 260 secondary teachers and found an overall low complexity, even though the problems met the curricular level.

10.2.3 Impact on Mathematics Education as a Discipline

The international LSA studies had different kinds of impact on the mathematics education community. As described in the previous section, LSA studies induced further research studies. However, there was and still is an additional impact on a meta-level, especially on the research content and research methods as well as on the discussion about the specific role of mathematics education research in the broader field of educational research.

Regarding the impact on the research content and methods, LSA studies firstly facilitated a normative discussion about the constructs which were addressed in LSA. Among others, there were scientific debates on the conceptualization of mathematical competence, the quality of mathematics instruction, and teacher professional competence. Related to these debates the goals of mathematics education in schools as well as in teacher education were scrutinized.

Secondly, LSA facilitated the "empirical turn" in mathematics education research. As an example, quantitative-empirical research articles in the main German journal on mathematics education (Journal für Mathematik-Didaktik; JMD) published after 2005 appear to be more sensitive to methodological challenges of empirical research than articles published in the 1980s and 1990s. Especially, decisions on adequate research designs, the quality of test items (item development based on

competence models) or the quality of processing raw data (reliability and objectivity) have become more important in the scientific discussion. In sum, the mathematics education community now has higher scientific standards concerning quantitative-empirical research.

A third important impact of LSA was and is that it raised questions about the features that constitute mathematics education as a scientific discipline. This question was triggered by the fact that the results on students' achievement in mathematics in TIMSS 1995 and PISA 2000 were mainly presented by researchers from educational science and psychology (although a few researchers from mathematics education were involved in these studies). Accordingly, the question came up what the specific contribution of mathematics educational research is in comparison to that of mathematics-related research in educational science. Interestingly, a similar question concerning the status of mathematics education came up again in the context of research on mathematics teachers' professional knowledge. For the conceptualization of the construct pedagogical content knowledge (PCK) and the development of related test items, researchers had to answer the question what mathematics educational knowledge should be. The challenge of this question became even clearer and was triggered by Lee Shulman's notion that PCK is simply an amalgam of content knowledge and pedagogical knowledge (Shulman 1987). This issue is still a matter of discussion within the educational community (Kirschner et al. 2017).

10.2.4 Impact on Schools and Mathematics Teachers

In addition to the impact of LSA on the scientific field of mathematics education, the LSA studies have consequences for the educational practice in schools and for mathematics teacher education at universities. The most striking consequence of LSA was the nationwide implementation of educational standards for mathematics combined with a monitoring system. In Germany, since 1945 the federal states are in charge of school education. The organization of the educational system in a decentralized manner was a consequence of the Nazi regime in Germany and was meant to avoid that a central government could control the educational system. Although a standing conference of all 16 ministries of education exists, until 2003 the federal states did not agree on precise goals and desired outcomes of school education (except for the German Abitur, which is the final examination from upper secondary schools and a general university-entrance qualification). In the 1990s, first ideas for educational standards in grade 10 were discussed. The results of TIMSS 1995 and PISA 2000 accelerated this process. In particular, the results of PISA 2000 in general and the huge differences in mathematics achievement between the different German federal states in particular put a lot of pressure on politicians. The German constitution (Grundgesetz) requires that the federal government and the governments of the federal states ensure an equal standard of living in all federal states (which means, in particular, equal educational opportunities). Hence, in 2003, educational standards for mathematics for the lower secondary level (grade 9/10) were implemented. In

2004, such standards were implemented for the primary level (grade 4), and in 2012 for the Abitur (upper secondary level, grade 12/13). In 2004, the federal "Institute for Educational Quality Improvement (IQB)" at the Humboldt University of Berlin was founded. The IQB is responsible for the evaluation of the educational standards. In 2006, a nationwide test-based monitoring system was implemented with high-quality test instruments. Teachers are obliged to administer these tests in grade 3 and grade 8, that is, one year before students should reach the educational standards for the primary or the secondary level, respectively. One important aspect of the monitoring system is that teachers analyze their students' test results so that they get information about their students' difficulties. Based on this information teachers can prepare specific learning support so that especially low achieving students get a chance to reach the educational standards after grade 4 or 10.

In addition to the above-mentioned changes of the educational system, new ideas were brought into teacher education and teacher professional development. For example, in 1998 the nationwide teachers' professional development (PD) program SINUS[1] started (and lasted until 2013). It included primary and secondary school teachers from more than 2500 schools. SINUS provided a specific infrastructure for PD (Prenzel et al. 2009). Teachers from different schools worked in local professional learning communities (PLC). The basis for each PLC was a pool of "state of the art material" prepared by researchers, which described topics from mathematics education. The PLC chose material from this pool and worked on the topics. Moreover, representatives of the PLCs met regularly in regional meetings and regional representatives met regularly in national meetings. The SINUS program for primary school teachers was evaluated based on TIMSS instruments. It turned out that students taught by teachers in so-called SINUS schools showed better achievement in mathematics than students taught in comparable schools without SINUS (Dalehefte et al. 2014).

There were also changes in the university-based teacher education system. Firstly, obligatory standards for mathematics teacher education in Germany were established in 2008. Secondly, the poor results of non-certified mathematics teachers in TEDS and the low mathematics achievement of their students in the evaluation of the educational standards had an important influence on the federal states to change the regulations for teacher education. In most federal states, all primary school teachers must now pass a substantial number of mathematics courses at the university. Moreover, all future teachers have to obtain a Master's degree or a comparable qualification before they are allowed to teach in school, regardless of school level (i.e., primary through upper secondary level).

In addition to the consequences on the institutional level described so far, LSA affected genuine mathematics educational research. We illustrate this impact using research on mathematical competence in the following two sections.

[1] SINUS = „Steigerung der Effizienz des mathematisch-naturwissenschaftlichen Unterrichts" (improving the efficiency of mathematics and science teaching).

10.3 Consequences from Large-Scale Assessment: An Example of Research on Mathematical Competence in Germany

In this section, we first explain why models of mathematical competence are necessary to implement and evaluate standards for school mathematics. We then describe how models of mathematical competence are developed in a process that takes into account both theories of mathematical development and empirical evidence. We illustrate this process using data from a large-scale national assessment in Germany. Finally, we show how additional qualitative analyses can contribute to our understanding of mathematical development on an individual level.

10.3.1 Understanding Mathematical Competence

As described above (Sect. 10.2) standards for school mathematics have been implemented in several countries in the last decades (e.g. Kultusministerkonferenz 2003, 2004, 2012, in Germany; National Council of Teachers of Mathematics 2000, in the U.S.). The influential Common Core State Standards Initiative (2012) asserts that standards "define what students should understand and be able to do" regarding mathematical content like number and quantity, algebra, functions, geometry, statistics and probability, as well as regarding typical mathematical practices like problem solving, reasoning and argumentation, modeling, use of tools, communication, use of structures and regularity. Standard-oriented classroom instruction aims at a profound understanding of mathematics and seeks to support students in applying their knowledge. Accordingly, standards are meant to support students' competence acquisition. This concept plays an important role with regard to standards, and it is also used in the German context. Weinert's (2001, pp. 27–28) definition is broadly accepted; it defines competences as „cognitive abilities and skills possessed by or able to be learned by individuals that enable them to solve particular problems, as well as the motivational, volitional and social readiness and capacity to utilize the solutions successfully and responsibly in variable situations." Standards thus refer to a conceptualization of competence that is similar to the concept of literacy on which the framework of PISA is based.

Students' competences may be assigned to mathematical content and practices but this assignment should take into account that students perform at different levels of proficiency. PISA 2012 provides a general description of these different levels based on international test data (OECD 2013). Table 10.1 summarizes six levels of proficiency of 15-year old students according to the OECD report (OECD 2013, p. 61).

This description of proficiency levels is important in order to understand students' performance particularly in the context of international comparisons. However, this description has limitations from a mathematics education point of view and with

Table 10.1 Levels of proficiency in mathematics in PISA (OECD 2013, p. 61)

Level	What students can typically do
6	At Level 6, students can conceptualise, generalise and utilise information based on their investigations and modelling of complex problem situations, and can use their knowledge in relatively non-standard contexts. They can link different information sources and representations and flexibly translate among them. Students at this level are capable of advanced mathematical thinking and reasoning. These students can apply this insight and understanding, along with a mastery of symbolic and formal mathematical operations and relationships, to develop new approaches and strategies for attacking novel situations. Students at this level can reflect on their actions, and can formulate and precisely communicate their actions and reflections regarding their findings, interpretations, arguments, and the appropriateness of these to the original situation
5	At Level 5, students can develop and work with models for complex situations, identifying constraints and specifying assumptions. They can select, compare, and evaluate appropriate problem-solving strategies for dealing with complex problems related to these models. Students at this level can work strategically using broad, well-developed thinking and reasoning skills, appropriate linked representations, symbolic and formal characterisations, and insight pertaining to these situations. They begin to reflect on their work and can formulate and communicate their interpretations and reasoning
4	At Level 4, students can work effectively with explicit models for complex concrete 3 situations that may involve constraints or call for making assumptions. They can select and integrate different representations, including symbolic, linking them directly to aspects of real-world situations. Students at this level can utilise their limited range of skills and can reason with some insight, in straightforward contexts. They can construct and communicate explanations and arguments based on their interpretations, arguments, and actions
3	At Level 3, students can execute clearly described procedures, including those that require sequential decisions. Their interpretations are sufficiently sound to be a base for building a simple model or for selecting and applying simple problem-solving strategies. Students at this level can interpret and use representations based on different information sources and reason directly from them. They typically show some ability to handle percentages, fractions and decimal numbers, and to work with proportional relationships. Their solutions reflect that they have engaged in basic interpretation and reasoning
2	At Level 2, students can interpret and recognise situations in contexts that require no more than direct inference. They can extract relevant information from a single source and make use of a single representational mode. Students at this level can employ basic algorithms, formulae, procedures, or conventions to solve problems involving whole numbers. They are capable of making literal interpretations of the results
1	At Level 1, students can answer questions involving familiar contexts where all relevant information is present and the questions are clearly defined. They are able to identify information and to carry out routine procedures according to direct instructions in explicit situations. They can perform actions that are almost always obvious and follow immediately from the given stimuli

regard to mathematics classroom practices. The reason is that the levels of proficiency are not "fine-grained" and do not specify details of mathematical processes and products. In particular, as this model is not meant to be a developmental model, it does not explain how learners may proceed from one level of proficiency to the next, or specify students' knowledge gaps when they fail to do so. Including such information in the model would be useful for analyzing student errors on specific items and for understanding learning processes (Vygotskij 1978).

Further refining the model would presuppose a better knowledge of learning processes, which would require detailed studies of students' performance. Moreover, profound theoretical considerations are necessary. It is important to identify the prerequisites for solving a specific problem with regard to the requirements concerning the mathematics behind the problem and with regard to general or everyday knowledge: For example, calculating 37 + 28 asks for basic numerical knowledge, and probably the application of various methods of calculation which presupposes knowing and understanding them. Determining the number of legs, which five dogs, eight birds, or seven goldfish have, on the other hand, requires basic numerical knowledge but also "every-day knowledge". Understanding mathematical competence must therefore take into account alternating empirical and theoretical considerations in order to properly describe competences and their development.

10.3.2 Modeling Mathematical Competence

Attempts have been made to describe mathematical competence and its development particularly from a theoretical point of view resulting in (mainly) normative models of mathematical competence (Reiss et al. 2007b). Evaluating these models has two aspects. On the one hand, we need empirical studies with an adequate number of tasks and students in order to verify or falsify the theoretical considerations. On the other hand, we need empirical studies, which demonstrate the distinction between levels of competence with regard to typical misunderstandings in order to initiate support in teaching these topics.

In the following, we illustrate how these aspects were addressed in a recent research agenda. First, we present data from a national assessment ("IQB-Ländervergleich") of a representative sample of 27,000 fourth-graders in Germany (Reiss et al. 2012), which aimed at monitoring the German educational standards in mathematics. Moreover, we describe how data may be used to inform teachers about strengths and weaknesses of their students.

10.3.2.1 Modeling Mathematical Competence with Data from a National Assessment

The test assessed mathematical competence in a variety of subdomains. Data were scaled such that the mean was 500 and the standard deviation 100. The breadth of the

levels was set to 70; however, with respect to low-performers and high-performers boundaries were open. In the following, we illustrate the model in the domain of numbers.

- Level I ($\leq$389): Basic technical background (routine procedures on the basis of simple conceptual knowledge)

Students at this level are familiar with simple mathematical terms and procedures. These terms and procedures can be reproduced correctly within a mathematical context, respectively within a context, which is familiar or well-trained. Specifically, students are proficient in exercises that require addition or multiplication of numbers up to ten. They can utilize these concepts during mental calculation exercises, partly written or written-only calculation exercises, if the exercises do not pose any specific difficulties. Additionally, they have to be able to apply these types of calculations in simple word problems correctly. Also, students can compare numbers based on their values and they can interpret numbers on place value panels without problems, especially in the range 1–1000.

- Level II (390–459): Simple use of basic knowledge (routine procedures within a clearly defined context)

Students are able to utilize their basic knowledge for simple, clearly structured and well-known mathematical problems. For example, problems requiring addition, subtraction, and multiplication can be solved using written or partly written algorithms. The students also conduct rough estimations and correctly recognize the dimensions of their results. In addition, students utilize the structure of the decimal system, and recognize general mathematical principles. These principles are taken into account when continuing simple number sequences, conducting structured counting procedures and engaging in systematic trials. They detect and apply familiar proportional attributes. Students can convert simple numbers into specified units, even when the units differ in value or when numbers have decimals. In addition, students can interpret clearly structured graphs, figures and tables even if they entail a large number of information.

- Level III (460–529 points): Recognition and utilization of relationships within a familiar context (both mathematical and factual)

Students are able to utilize their knowledge flexibly in various problems within a familiar context. Specifically, students handle numbers and operations within the curricular scope securely, and conduct numerical estimations well—even for large numbers. Students can recognize and describe structural aspects—at least when the contents have been practiced well. Students are able to continue number sequences that follow relatively complex rules; in addition, they identify incorrect numbers within straightforward number sequences.

- Level IV (530–599): Secure and flexible utilization of conceptual knowledge and procedures within the curricular scope

Students can utilize their mathematical knowledge securely, even when the context is not familiar. They correctly describe their own methods for calculation, understand and reflect on the approach of other students and are proficient in all mathematical calculations that are part of the curriculum. Following instructions, students manipulate and systematically change numbers in place value panels, even if the numbers are large (up to one million). They recognize the rules behind even complex number patterns and continue the patterns correctly. Calculations with quantities are performed securely and flexibly, especially calculations that involve approximations and estimations.

- Level V ($\geq$600): Modeling complex problems involving independent development of adequate strategies

Students can work on mathematical problems in each subject area adequately, securely and flexibly even if the context is unknown. Students are able to work on a highly advanced level both in using adequate strategies and in giving meaningful evaluations and generalizations. They use their in-depth knowledge from the curriculum flexibly even in unfamiliar situations, communicate their methods comprehensibly and explain why they chose their specific method. Mathematical arguments are assessed adequately, and complex situations are modeled and worked upon even if calculations require difficult processes such as the use of tables, compound quantities or numbers in decimal notation.

10.3.2.2 Modeling Competence Between Quantitative and Qualitative Approaches

Models of competence are based on well-defined demands; they identify mathematical tasks with varying difficulty and match students' competences and task characteristics. There is a straightforward scientific paradigm, the correct or wrong answer. However, such a paradigm has limitations because the correctness of answers does not allow immediate conclusions about the cognitive processes underlying these answers. In the following, we describe a study (for further details, see Obersteiner et al. 2015) wherein we tried to analyze students' answers in depth in order to better understand the differences between the various levels of competence with respect to errors and misconceptions. Data come from a sample of 3rd-graders who took part in an annual testing of their mathematical competences ("Vergleichsarbeiten"; VERA). The items addressed mathematical argumentation in the domain of number and operations as well as in the domain of patterns and structures.

Item 1 (see Fig. 10.1) addresses place values. The theoretical level of competence of this item according to the model was II, as was the empirical level. More than half of the students gave correct answers, 23% did not answer to the question. The most common mistake was referring to the ones only or to the tens only. These children obviously had a preliminary understanding of the place value system but were unable to extend it to this complex situation.

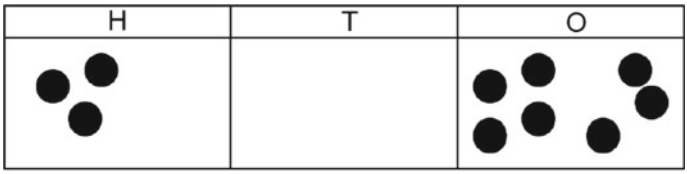

Fig. 10.1 Item 1: place value

Item 2 (see Fig. 10.2) addresses properties of numbers. In this case, the theoretical level of competence (III) was lower than the empirical level (IV), which means that the students had more problems with solving it correctly than was anticipated. About one third of the students gave correct answers, another third did not answer to the question. The most common mistake was not referring to the magnitude. Children were not able to use their knowledge of numbers in a problem-solving situation.

Item 3 (see Fig. 10.3) addresses number sequences. For this item, the theoretical and empirical levels of competence were identical and both high (IV for part a, V for part b). Only 27% (part a) and 14% (part b) of the students, respectively, answered the two questions correctly. The most common mistake was that students did not mention that the magnitudes of each two numbers add up to 100. Children's written work suggested that some were able to identify the correct solution but were unable to communicate their knowledge.

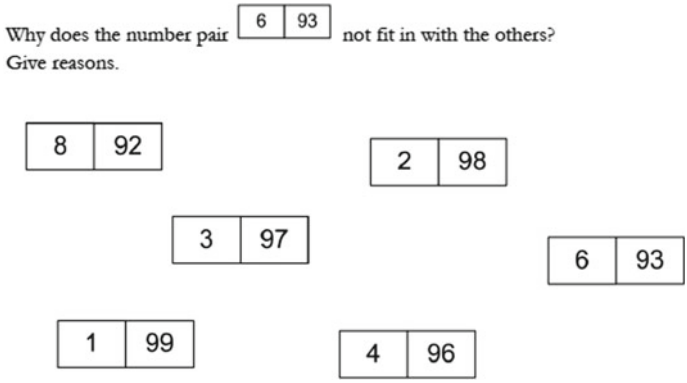

Fig. 10.2 Item 2: numbers pairs

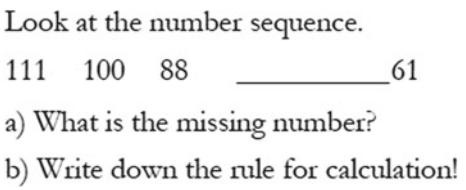

Fig. 10.3 Item 3: number sequences

10.3.3 Summary

We have illustrated how international large-scale assessments and the introduction of educational standards for school mathematics had an impact on mathematics education research and practice in Germany. Competence models have been developed that were based on both theoretical considerations and empirical evidence from large-scale assessments. To increase our understanding of how we can support students develop from a lower level of competence to a higher one, additional qualitative approaches that include an analysis of where exactly students go wrong at a certain level of competence can be fruitful. Such approaches can eventually lead to more fine-grained models of mathematical competence and its development.

10.4 Consequences from Large-Scale Assessment: Competence Models and Evaluation of Classroom Practice in Austria

The current system of LSA in Austria comprises three strands: international assessment as external benchmark, standardized assessment of national education standards (BIST-Ü) as a main lever for school quality development, and informal competence measurement (IKM) as a self-evaluation tool for teachers. This section first provides a brief history of the system and then focuses on the two strands of national assessments. We introduce the competence models and the national education standards, describe briefly the implementation of the assessment, and then explain the reporting system.

10.4.1 Development of Large-Scale Assessment in Austria

As in other German speaking countries, LSA are a relatively young phenomenon in Austria. By participating in TIMSS in 1995, Austria, like Germany (see Sect. 10.2), started to participate in international LSAs. The results were mixed: Austria was in the group of top performers on the elementary school level but not on the lower secondary school level. Austria had a very small group of top performers, and there was a

comparably large gender gap (Beaton et al. 1997; Mullis et al. 1997). However, the results were not of much interest to policy makers or the broader public. This changed in 2001, when the results from PISA 2000 were presented and publicly discussed. Of most concern was (and is) the group of students identified as "students at risk". In mathematics, 22% of students in Austria currently fall into this category (Suchán and Breit 2016). Moreover, there is a comparably small group of high performing students.

In response to these results, the Ministry of Education of Austria assigned a committee to devise a "master plan". One of the goals of the committee was to change the culture in the school system from input-oriented to output-oriented. In the report (Eder et al. 2005), the committee listed a number of measures intended to augment school quality, amongst others the introduction of national educational standards and the assessment thereof. At about the same time, competence models for the subject domains were developed and empirically validated in pilot studies. In 2009, a law was passed (Bundesgesetzblatt II Nr. 1/2009) that defined national educational standards in mathematics, German (both grades 4 and 8), and English (grade 8 only), which were based on national curricula. The law also mandated regular assessment of the competences and "meaningful feedback" to students, teachers, and school principals as well as a system of monitoring reports in order to foster evidence-driven activities. A baseline study was conducted in 2009 (lower secondary school level, n = 204 schools) and 2010 (elementary school level, n = 267 schools) in order to measure the level of competences before national education standards were implemented. Since 2012, competences in mathematics, German, and English are assessed, alternating every year in a census in grades 4 and 8. Feedback is prepared by the BIFIE on several levels for each national LSA (see Table 10.2).

10.4.2 Mathematics Competence Models in Austria

For all subjects and grades covered, competence models were devised. In 2007, the Institute for Didactics of Mathematics at the University of Klagenfurt (Austria) pub-

Table 10.2 Feedback recipients

Recipients	Format	N
Students	Online, individualized	About 76',000 per grade
Teachers (classes, groups)	Online, for each class taught	About 4000 in lower secondary/6000 in elementary
Schools	Online, per school and class	About 1400 in lower secondary/3000 in elementary
School authorities	Online and in print	99 school districts
Regional (federal state)	Online and in print	9 states
National	Online and in print	1 report

lished a brochure in which a competence model for mathematics in lower secondary schools was introduced (Institut für Didaktik der Mathematik 2007). In this section, we only describe the model used for grade 8 mathematics. The competence model uses three dimensions to describe mathematical competence: process, content, and level of complexity (see Table 10.3).

The process domains as well as the content domains contain four distinct areas each; the level of complexity is divided into three hierarchical levels. Since these dimensions of content, processes, and complexity are orthogonal to each other (see Fig. 10.4), the model consists of $4 \times 4 \times 3 = 48$ nodes altogether. For every node, a can-do statement describes students' skills and abilities (Verordnung der Bundesministerin für Unterricht, Kunst und Kultur über Bildungsstandards im Schulwesen 2009), for example "the students are able to describe algebraic, tabular or graphic representations of structures and (functional) relations and interpret both in a given context. In doing so, it is necessary to make connections to other mathematical contents (terms, theorems, representations) or processes" (process 3/content 2/complexity level 2, see grey cube in Fig. 10.4).

Mathematics materials used in school are supposed to use this structure and cover all process and content domains. Test items for national tests are developed on the basis of these nodes; only items that can be classified unambiguously on a node are used in the test, in the sense that the group of test developers and at least one external expert reach consensus on the classification. Since the level of complexity is not equal to empirical difficulty, at reporting stage, empirical difficulty along with the classification according to educational standards (competence level) are also published. As an example, the item in Fig. 10.5 has been classified as follows.

- Process domain: demonstrating, modeling
- Content domain: statistical representation, measures of central tendency and variance
- Competence level: 2 (indicating standards reached, see Table 10.3)

Table 10.3 Dimensions of mathematical competence

Process domains (Handlungsbereiche)	Content domains (Inhaltsbereiche)	Level of complexity (Komplexitätsbereiche)
P1: demonstrating, modeling	C1: numbers and units	L1: use of basic skills and knowledge
P2: calculating, operating	C2: variable, functional dependency	L2: build connections/connect
P3: interpreting	C3: geometric figures and shapes	L3: reflect, use knowledge of reflection
P4: explaining, reasoning	C4: statistical representations, measures of central tendency and variance	

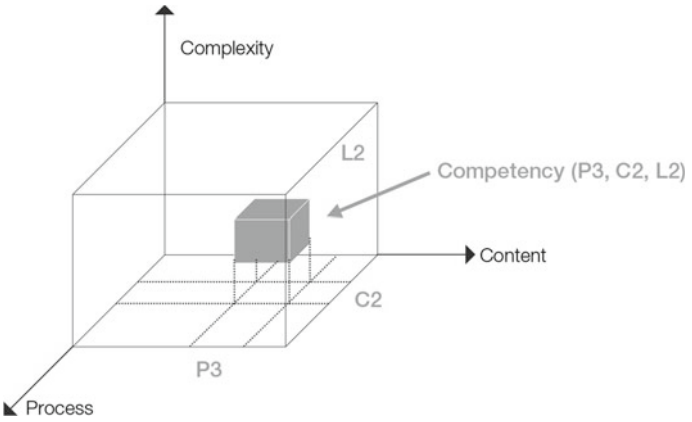

Fig. 10.4 Competence model, mathematics, grade 8. Adapted from Institut für Didaktik der Mathematik (2007)

Example item (translated): Alina and Christoph want to go on a 5-day canoe trip. They plan to do 15 km per day on average. After 4 days, they have managed to do the following distances:

Day	1	2	3	4	5
Distance	17	12	14	16	?

How many kilometers do they have to do on day 5 to reach the average of 15 km per day?
Write your answer in the box.

Fig. 10.5 Example item

- Answer: 16 km. Empirical difficulty: 53% of students in Austria solved this item (BIFIE 2012a).

An expert group consisting of mathematics teachers and mathematics education researchers described in four levels the degree to which the education standards are reached (see Table 10.4).

In contrast to the descriptions used in international assessments, these levels were defined and described using a Standard Setting method called Item-Descriptor-Matching method (Freunberger 2013). The method requires comparison of the performance shown on the test items with the theoretical framework (in this case, the national curriculum and national standards). The exercise was undertaken before the main assessment was carried out, using field trial data to sort items by empirical difficulty. As the first national assessment in mathematics was strongly debated, this was necessary to keep the Standard Setting out of political discussion. The performance level descriptors were centered around level 2, "standards reached", as this represents the desired outcome. The competence model and the performance level

Table 10.4 Levels of competence, mathematics, grade 8

Level 3—standards exceeded
Students possess fundamental knowledge and skills in all parts of the mathematics curriculum and advanced knowledge structures, which exceed the requirements of level 2, specifically more pronounced abilities of abstraction and higher proficiency in combining parts of knowledge, methods or rules. They are able to apply these independently in novel situations in a flexible way
Level 2—standards reached
Students possess fundamental knowledge and skills in all parts of the mathematics curriculum and can use these in a flexible way. They are able to find and apply problem solving strategies, to describe and reason about their approach. They are able to handle verbal, graphical and formal representations of mathematical facts in a flexible way and can apply these appropriately. They are able to extract relevant information from differently represented facts (e.g. texts, data material, graphics) and can interpret them in the respective context. They are able to relate their mathematical knowledge and can check, evaluate and/or reason about mathematical statements
Level 1—standards partially reached
Students possess fundamental knowledge and skills in all parts of the mathematics curriculum and can master reproductive tasks and carry out routine procedures
Below level 1—standards not reached
There is no further description of this level, as students on this level typically show insufficient competences/skills

descriptors represent the core of the feedback to all stake-holders and can be found in all reports (see Table 10.2).

10.4.3 Reporting of Assessment of National Education Standards

10.4.3.1 Reporting on the System Level

System monitoring is one of the main goals of the assessment of national education standards. A detailed report is published every year on a national and a regional level (cf. Schreiner and Breit 2012a, b; Schreiner et al. 2018). The outcomes should inform education policy makers about issues and areas for improvement. Since the assessment has only started the second cycle, it is yet too early to identify sustainable development. However, we can state that students have performed better in mathematics in 2017 (national average 542 points) than in 2012 (535 points) and in the baseline assessment in 2009 (500 points). This increase in performance was highest in the content domain of statistics (550 points vs. 544 points vs. 500 points). In part, this might have to do with instructional practices; teachers can choose not to cover all areas of the curriculum. Presumably, statistics had been covered by teachers less often than other areas prior to the assessments: A study among first-year students of mathematics education showed that statistics is also the most unpopular content

area, and that students complained about "comprehension difficulties", "bad lessons" and "no relation to everyday life" (Süss-Stepancik and George 2016). This probably also affects teaching (Süss-Stepancik and George 2017). In the years preceding more recent assessments, the (dis-)likes for content areas and process domains could have been evened out by the fact that all areas are covered by the national tests because teachers might say "now that you test it, we teach it". Teachers may also be more aware of the assessments because instructional material now contains more references to national education standards and competence models than in 2009.

10.4.3.2 Reporting on the School and Class Level

The feedback to school principals and teachers is designed to foster school development and improvement of instructional practice. Based on a model of factors influencing development of school quality and teaching strategies (Helmke 2004), Wiesner et al. (2015) devised a framework modeling the use of feedback for schools and teachers for development. A central point is the standardized assessment which gives objective feedback, in how far schools and classes are able to convey the competences described in the national education standards. The first step comprises the correct analysis and interpretation of the results in order to identify weaknesses and potentials for improvement. Reflecting the outcomes, teachers and school heads then can design actions for improvement. In an evaluation step, the impact of the implemented actions is scrutinized and the quality development cycle starts again. In order to support analysis and interpretation of results, individualized feedback for each class taught is created for to teachers using standard templates (see Table 10.2). This feedback describes not only the outcomes but also other variables like students' socio-economic background, their wellbeing at school, and their motivational situation. Moreover, it includes a "fair comparison" (Pham and Robitzsch 2014), that is, a comparison with schools/classes showing similar characteristics. In addition to feedback on levels of mathematics competence and feedback in terms of points reached overall as well as feedback on the results of subgroups in the class, teachers get feedback on the results of their class(es) in the process and content domains. In contrast to conventions in many other countries, school or class results are not reported publicly as such reporting is not regarded as helpful in terms of quality development. In order to give schools space to develop and work with the results, no school ranking is published. Aggregated reports on regional and national level are published online without reference to specific schools. School results are reported to schools and school authorities only.

As an example, Fig. 10.6 shows the feedback on the process domains (note: data are fictitious). It contains the Austrian mean for all domains (see flag) as well as the school mean (see punctuated line) and the class mean (dot with confidence interval). Individual students are represented as dots. The higher up the bar, the higher the achievement; the wider spread the dots are, the larger the variance within the class. The table below the chart lists the number of students in school and class and the mean including the confidence interval for each domain.

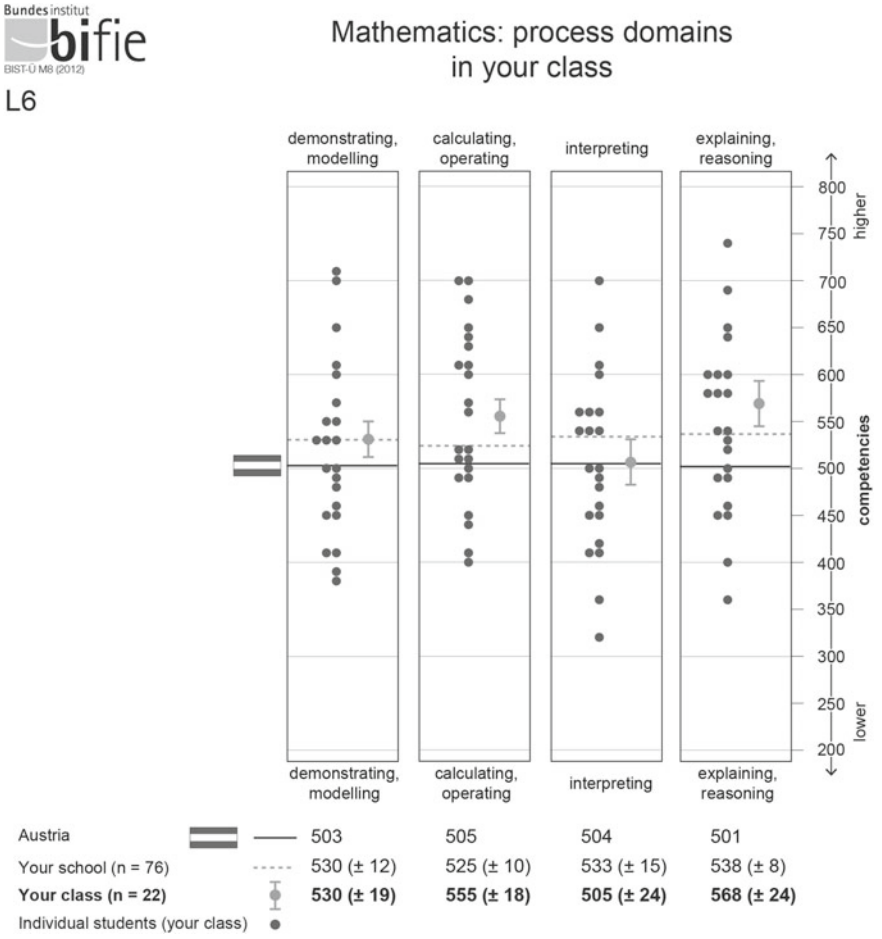

Fig. 10.6 Results on class (group) level by process domain (BIFIE 2012b)

10.4.3.3 Other Report Levels

School authorities receive school results only, without details on the classroom level. In many cases, this level of granularity is acceptable as in Austria the student population within schools tends to be more homogenous than between schools. Therefore, classes within the same school tend to perform more consistently than classes in different schools. By law, students are entitled to their individual feedback. During the test session, they receive a unique code for their individualized feedback available via a secure online platform. Although students are the biggest recipient group, the impact of the feedback on them is probably quite negligible as the results must not be used for grading. The student data is considered confidential and no-one but the student should have access to it.

10.4.4 Informal Competence Measurement

In contrast to the assessment of national education standards, the informal competence measurement is focusing on the classroom level only (Wiesner et al 2017). Teachers can decide whether or not to use this free, web-based tool in their classes. The online platform can score selected response items automatically. Teachers only need to code open response items themselves. The platform instantly generates feedback on the class level and on the level of individual students. It provides feedback for mathematics in general and for process and content domains separately. For each item, national benchmarks stemming from pilot tests are provided. Since the use of the tool is voluntary, the exposure is somewhat hard to measure; roughly 15% of students take it in grade 6 and about 25% of students in grade 7 (in mathematics). Teachers may choose to share the results but are not required to.

Table 10.5 Comparison of mandatory national assessment and informal competence measurement

	Assessment of national education standards	Informal competence measurement
Status	• Mandatory	• Voluntary (mostly, teachers decide use)
Grade(s)	• 8	• 6 and 7 (grade 5 in preparation, grade 8 in planning)
Cycle	• Every 5 years	• Every year
Reach	• Nearly all students in grade 8 (about 76,000) • 4000 classes in 1400 schools	• Between 15 and 25% of students in around 40% of schools (mathematics only)
Administration	• Highly standardized • Trained test administrators carry out test • Paper-pencil tests only	• Somewhat standardized (e.g., timing of test, test materials) • Subject teachers test their students • Online test administration
Scoring	• Materials are collected and are scored centrally • Items in open ended format are coded by trained coders using standardized coding guide	• Selected response items are scored automatically • Open ended format-items are scored by class teachers; use of standardized coding guide is suggested
Analysis methods	• IRT models	• Percent correct
Feedback	• Feedback is sent to schools and teachers months later, students are anonymous • On aggregated levels, school authorities and public are informed	• Online tool generates feedback immediately on class and student level. Teachers can identify individual students • Access to feedback is restricted to teachers (self-evaluation)

(continued)

Table 10.5 (continued)

	Assessment of national education standards	Informal competence measurement
Use	• Teachers, school heads: Birds-eye view on educational output, possibility to identify weaknesses/strengths • Comparison with other schools and classes in similar settings ("fair comparison") • School authorities: aggregated feedback on school level • System monitoring • Development of education quality through developments at schools, teacher education, education policy, curricular developments	• Teachers: overview of group as well as diagnosis of individual students' achievement levels • Results must not be used for grading • No other users: no reports for school head, school authorities, system monitoring. Teachers may decide to share the results • Support of individual students by teachers • Supporting teachers through focused and detailed feedback about class and individuals, in relation to education standards
Contextual information	• Context questionnaires on student and school level allow capture of students and school characteristics	• Minimal context questions (gender, migration background); restricted to pilot test

The competence models used in this tool are the same as the models that are used for the national assessment. The comparison in Table 10.5 is restricted to lower secondary school.

10.4.5 Summary

The different strands of large-scale assessment in Austria allow for different viewpoints on students' mathematics competence, depending on the main uses and users. The international LSAs show how students in a specific country perform compared to other countries' students. This is useful feedback for policy makers and interesting for the public. The national assessments allow for comparable and meaningful feedback to teachers and school principals as well as students and parents, school authorities and the broader public about the levels of competence reached as well as areas where there is room for improvement. Figure 10.7 summarizes the intended interactions between curriculum, competency models, national education standards, assessments, and feedback.

As illustrated in this figure, there are many channels for feedback to relevant actors and stakeholders in the system. Further evaluation of the impact of national education standards and their standardized and mandatory assessment will hopefully further clarify the role of education standards.

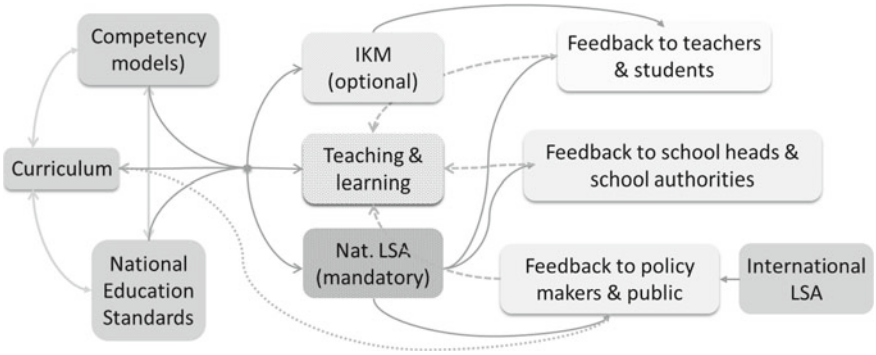

Fig. 10.7 Interactions and embedding of large-scale assessments in Austria

10.5 Summary and Discussion

Large-scale assessment plays an increasingly important role on various levels of educational research and practice in German-speaking countries. The developments in Germany and Austria described in this chapter show the potential of LSA in several respects. Large-scale studies have an impact on curriculum development, policy making, and the development of theories of mathematical competence. They also can inspire further research to address issues that large-scale studies have identified as being relevant for competence development. For example, the discussion about what mathematical competence should be has led to the agreement that competence is more than proficiency and encompasses cognitive but also non-cognitive aspects, such as motivation and interest (Weinert 2001). Also, it would be useful if competence models provided information not only on students' competence but also on what students did not yet understand. Such information would be particularly valuable for teachers who need information on students' strengths and weaknesses in order to support their learning. Thus, future research could focus on these aspects.

However, large-scale assessments also have limitations. In the remainder of this section, we first reflect on general issues of large-scale assessments, and then discuss specific limitations and potential risks for research and practice.

10.5.1 General Issues of Large-Scale Assessments from an International Perspective

We discuss general issues of large-scale assessments from three perspectives: their purposes, their processes, and their outcomes.

10.5.1.1 Purposes: The Studies Reflect the Needs of a Country

The decision to conduct large-scale studies is often accompanied by the intention to solve specific educational problems in a country. In the case of Germany and Austria, for example, the starting point of large-scale studies was mainly influenced by the international assessments of students' mathematics performances (i.e., TIMSS and PISA). The results of those studies were disappointing and received much public attention, which in turn pushed stakeholders to improve the situation. Empirical studies provided information about the situation in classrooms. In other words, they enabled investigating certain phenomena more closely to eventually improve the situation for teachers and students, and for developing the curriculum.

In Germany, to support primary and secondary school teachers' professional development, the country started a nationwide teacher professional development program (SINUS). Moreover, there were studies investigating teachers' competencies in teaching mathematics from a national perspective (e.g., COACTIV), as well as from an international perspective (e.g., TEDS-M). In a next step, studies investigated not only school students' and teachers' mathematical competences but also those of adults (e.g., PIAAC). In order to trace the development of competence, the NEPS project was started in 2008. It includes participants of a wide age range to assess their mathematics competence both cross-sectionally and longitudinally.

10.5.1.2 Processes: Revealing the Functions of Studies

To discuss issues related to the aspect of processes, we make an analogy to the TIMSS framework (see Mullis and Martin 2013), which describes how OECD adapts the processes of knowing, applying, and reasoning in designing test items and assessing what students learned in mathematics. We consider these three processes suitable for interpreting how the countries view large-scale studies through the process of (1) knowing what phenomenon they want to investigate, and (2) applying what they have known from precedent studies to explore the new domains they want to investigate further (i.e., a new round of knowing process), or to solve the present educational problems they met. In order to process the application, (3) re-analyzing the collected data with new methods creates opportunities for new studies exploring additional phenomena or outcomes.

The primary data collected from large-scale assessments provide abundant and useful information. The information from these primary data usually contributes directly to educational fields, such as policy making and comparison within a country or between countries. The other potential method to deal with the abundant data is secondary analysis (Glass 1976). Glass introduces the two purposes of this method, one is the re-analysis of data for the purpose of answering the original research question with better statistical techniques, and the other is to answer new questions with old data. An example is the secondary analysis of TIMSS data (Robitaille and Beaton 2002).

In the example of research on mathematical competence in Germany, the purpose was for example the assessment of fourth-graders' mathematical competence, aiming at developing a competence model. This model is related to competence levels of students' performances in PISA 2012 (OECD 2013). The primary data were collected from the survey of nationwide students with validated items. After analyzing those primary data, the competence model was constructed for specific mathematical content areas. In a next step, students' written responses were investigated to better understand students' errors and misconceptions (see Sect. 10.3).

In Austria, the process was similar to Germany, but the example of Austria shows how competence models can be useful to provide feedback on different levels of the educational system. Data with different ecologies were used to identify new educational phenomena in different schools or classes (King 1997). In this way, large-scale studies can inform policy-makers to make realistic decisions based on empirical data.

10.5.1.3 Outcomes of Large-Scale Studies

As already mentioned, the purpose of large-scale studies is often to solve specific educational problems. The examples from Germany and Austria reported in this chapter showed how the development of competence models benefited from students' performance data from large-scale assessments. Such competence models are dynamic regarding their contents and their participants in related assessments. Moreover, the Austrian study further applied the survey of competence into the system monitoring and suggestions to instructional improvement as the intended outcomes. The example of Germany illustrated how data can be used for another outcome, namely secondary analyses of students' answers to better understand their reasoning and eventually inform instruction concretely and constructively.

10.5.2 Potential Risks for Research and Practice

There are potential risks of LSA for research and practice. Although LSA induced progress and changes in mathematics educational research as well as in mathematics classroom, there are some aspects which can be considered as critical.

A first aspect is that LSAs provide a huge amount of empirical data on students' and teachers' mathematics achievement. This can create an imbalance between empirical and theoretical perspectives because it can be tempting to mix up empirical data with empirical evidence. Hence, it is important to keep in mind that generating and interpreting empirical data requires theories, and that theories are also necessary to transform empirical data into empirical evidence. This is particularly important in cases when LSA data develop a "life on their own". Examples are the increasing amounts of purely descriptive empirical results with little scientific impact (i.e. low

explanatory power) or the illusion of "exact" results due to empirical data-bases and over-interpretation of statistical results.

A second aspect is the norm setting. For example,

- in the research context, we can observe the situation that LSA becomes self-referential in the sense that LSA becomes an own research topic,
- in the context of educational goals, we can observe that benchmarks for learning outcomes are determined by LSA data instead of normative educational goals,
- in the context of educational policy, we can observe that LSA results promote pragmatic decisions and substitute educational visions as guiding principles.

In conclusion, the research community should critically analyze which role LSA already has and which role LSA should have for research and practice in mathematics education.

References

Bauersfeld, H. (1978). Kommunikationsmuster im Mathematikunterricht. Eine Analyse am Beispiel der Handlungsverengung durch Antworterwartung. In H. Bauersfeld (Ed.), *Fallstudien und Analysen zum Mathematikunterricht* (S. 158–170). Hannover: Schroedel.

Baumert, J., Klieme, E., Neubrand, M., Prenzel, M., Schiefele, U., Schneider, W., et al. (Eds.). (2001). *PISA 2000. Basiskompetenzen von Schülerinnen und Schülern im internationalen Vergleich*. Opladen: Leske + Budrich.

Baumert, J., Lehmann, R., Lehrke, M., Schmitz, B., Clausen, M., Hosenfeld, I., et al. (1997). TIMSS – Mathematisch-naturwissenschaftlicher Unterricht im internationalen Vergleich. Deskriptive Befunde. Opladen: Leske + Budrich.

Beaton, A. E., Mullis, I. V. S., Martin, M. O., Gonzalez, E. J., Kelly, D. L., & Smith, T. A. (1997). *Mathematics achievement in the middle school years: IEA's third international mathematics and science study (TIMSS).* Chestnut Hill: Boston College.

BIFIE. (2012a). Beispielitems aus der Standardüberprüfung Mathematik 2012 für die 8. *Schulstufe*. https://www.bifie.at/system/files/dl/Beispielaufgaben_BIST-UE-M8-2012.pdf. Accessed March 14, 2017.

BIFIE. (2012b). Rückmeldung an die Lehrer/innen. *Standardüberprüfung M8—2012*. https://www.bifie.at/node/1689. Accessed March 14, 2017.

Blömeke, S., Kaiser, G., & Lehmann, R. (Eds.). (2010). *TEDS-M 2008 - Professionelle Kompetenz und Lerngelegenheiten angehender Primarstufenlehrkräfte im internationalen Vergleich*. Münster: Waxmann.

Blum, W., & Neubrand, M. (Eds.). (1998). *TIMSS und der Mathematikunterricht. Informa-tionen, Analysen und Konsequenzen*. Hannover: Schroedel.

Common Core State Standards Initiative. (2012). *Common core state standards: Mathematics*. http://www.corestandards.org/Math. Accessed March 30, 2017.

Dalehefte, I. M., Wendt, H., Köller, O., Wagner, H., Pietsch, M., Döring, B., et al. (2014). Bilanz von neun Jahren SINUS an Grundschulen in Deutschland. Evaluation der mathematikbezogenen Daten im Rahmen von TIMSS 2011. *Zeitschrift für Pädagogik, 2*(60), S.245–S.263.

Deutsches PISA-Konsortium. (2001). *PISA 2000: Basiskompetenzen von Schülerinnen und Schülern im internationalen Vergleich*. Opladen: Leske + Budrich.

Duchhardt, C., Jordan, A.-K., & Ehmke, T. (2017). Adults' use of mathematics and its influence on mathematical competence. *International Journal of Science and Mathematics Education, 15*(1), 155–174.

Eder, F., Haider, G., Specht, W., Spiel, C., & Wimmer, M. (2005). zukunft:schule. Strategien und Maßnahmen zur Qualitätsentwicklung. *Abschlussbericht der Zukunftskommission*. https://www.bifie.at/system/files/dl/Reformkonzept%20zukunft%20schule%20II%20Abschlussbericht%202005.pdf. Accessed March 14, 2017.

Freunberger, R. (2013). Standard-Setting Mathematik 8. Schulstufe. *Technischer Bericht*. https://www.bifie.at/wp-content/uploads/2017/05/StaSett_M8_TechReport_sV__2013-05-15.pdf. Accessed May 17, 2018.

Glass, G. V. (1976). Primary, secondary, and meta-analysis of research. *Educational Researcher, 5*(10), 3–8.

Helmke, A. (2004). Von der Evaluation zur Innovation: Pädagogische Nutzbarmachung von Vergleichsarbeiten in der Grundschule. *SEMINAR – Lehrerbildung und Schule, 2*, 90–112.

Institut für Didaktik der Mathematik. (Eds.). (2007). Standards für die mathematischen Fähigkeiten österreichischer Schülerinnen und Schüler am Ende der 8. *Schulstufe.* https://www.uni-klu.ac.at/idm/downloads/Standardkonzept_Version_4-07.pdf. Accessed March 14, 2017.

Jordan, A., Krauss, S., Löwen, K., Blum, W., Neubrand, M., Brunner, M., et al. (2008). Aufgaben im COACTIV-Projekt: Zeugnisse des kognitiven Aktivierungspotentials im deutschen Mathematikunterrichts. *Journal fürMathematikdidaktik, 29*(2), 83–107.

King, G. (1997). *A solution to the ecological inference problem: Reconstructing individual behavior from aggregate data*. Princeton, NJ: Princeton University Press.

Kirschner, P. A., Verschaffel, L., Star, J., & Van Dooren, W. (2017). There is more variation within than across domains: An interview with Paul A. Kirschner about applying cognitive psychology-based instructional design principles in mathematics teaching and learning. *ZDM–Mathematics Education, 49*(4), 637–643.

Komorek, E., Bruder, R., Collet, C., & Schmitz, B. (2007). Contents and results of an intervention in maths lessons in secondary level I with a teaching concept to support mathematic problem-solving and self-regulative competencies. In M. Prenzel (Ed.), *Studies on the educational quality of schools. The final report on the DFG Priority Programme* (pp. 175–196). Münster: Waxmann.

Kultusministerkonferenz. (2003). *Bildungsstandards im Fach Mathematik für den mittleren Schulabschluss*. Bonn: KMK.

Kultusministerkonferenz. (2004). *Bildungsstandards im Fach Mathematik für den Primarbereich*. Bonn: KMK.

Kultusministerkonferenz. (2012). *Bildungsstandards im Fach Mathematik für die Allgemeine Hochschulreife*. Bonn: KMK.

Kunter, M., Baumert, J., Blum, W., Klusmann, U., Krauss, S., & Neubrand, M. (2013). *Cognitive activation in the mathematics classroom and professional competence of teachers. Results from the COACTIV project*. New York, NY: Springer.

Lipowsky, F., Rakoczy, K., Drollinger-Vetter, B., Klieme, E., Reusser, K., & Pauli, C. (2009). Quality of geometry instruction and its short-term impact on students? Understanding of pythagorean theorem. *Learning and Instruction, 19*(6), 527–537.

Mullis, I. V. S., & Martin, M. O. (Eds.). (2013). *TIMSS 2015 assessment frameworks*. Chestnut Hill, MA: TIMSS & PIRLS International Study Center, Lynch School of Education, Boston College.

Mullis, I. V. S., Martin, M. O., Beaton, A. E., Gonzalez, E. J., Kelly, D. L., & Smith, T. A. (1997). *The mathematics achievement in the primary school years: IEA's third international mathematics and science report*. Chestnut Hill: Boston College.

National Council of Teachers of Mathematics. (2000). *Principles and standards for school mathematics*. Reston, VA: National Council of Teachers of Mathematics.

Neubrand, J. (2002). *Eine Klassifikation mathematischer Aufgaben zur Analyse von Unterrichtssituationen – Selbsttätiges Arbeiten in Schülerarbeitsphasen in den Stunden der TIMSS Video-Studie*. Hildesheim: Franzbecker.

Neumann, I., Duchhardt, C., Grüßing, M., Heinze, A., Knopp, E., & Ehmke, T. (2013). Modeling and assessing mathematical competence over the lifespan. *Journal of Educational Research Online, 5*(2), 80–109.

Obersteiner, A., Moll, G., Reiss, K., & Pant, H. A. (2015). Whole number arithmetic—Competency models and individual development. In X. Sun, B. Kaur, & J. Novotná (Eds.), *Proceedings of the 23rd ICMI Study Conference: Primary Mathematics Study on Whole Numbers* (pp. 235–242). Macao, China: University of Macau.

OECD. (2013). *PISA 2012 results: What students know and can do: Student performance in mathematics, reading and science* (Vol. I). Paris: OECD.

Pham, G., & Robitzsch, A. (2014). „Fairer Vergleich" [„fair comparison"]. Technische Dokumentation – BIST-Ü Englisch, 8. Schulstufe, 2013. BIFIE: Salzburg. https://www.bifie.at/system/files/dl/TD_Fairer_Vergleich_E8.pdf. Accessed March 13, 2017.

Prenzel, M., Stadler, M., Friedrich, A., Knickmeier, K., & Ostermeier, C. (2009). *Increasing the efficiency of mathematics and science instruction (SINUS): A large scale teacher professional development programme in Germany*. Kiel: Leibniz-Institute for Science Education (IPN). Retrieved from https://www.ntnu.no/wiki/download/attachments/8324749/SINUS_en_fin.pdf.

Rammstedt, B. (2013). *Grundlegende Kompetenzen Erwachsener im internationalen Vergleich Ergebnisse von PIAAC 2012*. Münster: Waxmann.

Reiss, K., Heinze, A., Kessler, S., Rudolph-Albert, F., & Renkl, A. (2007). Fostering argumentation and proof competencies in the mathematics classroom. In M. Prenzel (Ed.), *Studies on the educational quality of schools. The final report on the DFG Priority Programme* (pp. 251–264). Münster: Waxmann.

Reiss, K., Heinze, A., & Pekrun, R. (2007). Mathematische Kompetenz und ihre Entwicklung in der Grundschule. In M. Prenzel, I. Gogolin, & H. H. Krüger (Eds.), *Kompetenzdiagnostik. Sonderheft 8 der Zeitschrift für Erziehungswissenschaft* (S. 107–S. 127). Wiesbaden: Verlag für Sozialwissenschaften.

Reiss, K., Roppelt, A., Haag, N., Pant, H. A., & Köller, O. (2012). Kompetenzstufenmodelle im Fach Mathematik. In P. Stanat, H. A. Pant, K. Böhme, & D. Richter (Eds.), *Kompetenzen von Schülerinnen und Schülern am Ende der vierten Jahrgangsstufe in den Fächern Deutsch und Mathematik. Ergebnisse des IQB-Ländervergleichs 2011* (S. 72–S. 84). Münster: Waxmann.

Robitaille, D., & Beaton, A. E. (Eds.). (2002). *Secondary analysis of the TIMSS data*. Dordrecht: Kluwer Academic Publishers.

Schreiner, C., & Breit, S. (2012a). Standardüberprüfung 2012. Mathematik, 8. Schulstufe. *Bundesergebnisbericht*. https://www.bifie.at/system/files/dl/01_BiSt-UE_M8_2012_Bundesergebnisbericht.pdf. Accessed March 14, 2017.

Schreiner, C., & Breit, S. (2012b). Standardüberprüfung 2012. Mathematik, 8. Schulstufe. *Landesergebnisberichte*. https://www.bifie.at/node/1949. Accessed March 14, 2017.

Schreiner, C., Breit, S., Pointinger, M., Pacher, K., Neubacher, M., & Wiesner, C. (Eds.). (2018). Standardüberprüfung 2017. Mathematik, 8. Schulstufe. *Bundesergebnisbericht*. https://www.bifie.at/wp-content/uploads/2018/02/BiSt_UE_M8_2017_Bundesergebnisbericht.pdf. Accessed May 17, 2018.

Shulman, L. S. (1987). Knowledge and teaching: Foundations of the new reform. *Harvard Educational Review, 57,* 1–22.

Stigler, J. W., Gonzales, P., Kawanaka, T., Knoll, S., & Serrano, A. (1999). *The TIMSS videotape classroom study. Methods and findings from an exploratory research project on eighth-grade mathematics instruction in Germany, Japan, and the United States*. Washington, D.C.: U.S. Department of Education.

Suchán, B., & Breit, S. (Eds.). (2016). *PISA 2015. Grundkompetenzen am Ende der Pflichtschulzeit im internationalen Vergleich*. Graz: Leykam.

Süss-Stepancik, E., & George, A. C. (2016). Was ich an Mathe mag – oder auch nicht! Epistemologische Beliefs von Studienanfängern/anfängerinnen an der PH NÖ. *Open Online Journal for Research and Education, 5*(1), 50–62.

Süss-Stepancik, E., & George, A. C. (2017). Einstellungen von Mathematik-Lehrenden und Lehramtsstudierenden zu mathematischen Teilkompetenzen. In Institut für Mathematik der Universität Potsdam (Ed.), *Beiträge zum Mathematikunterricht 2017*. Münster: WTM-Verlag.

Verordnung der Bundesministerin für Unterricht, Kunst und Kultur über Bildungsstandards im Schulwesen. BGBl. II Nr. 1/2009. https://www.ris.bka.gv.at/Dokument.wxe?Abfrage=BgblAuth&Dokumentnummer=BGBLA_2009_II_1. Accessed March 14, 2017.

Vygotskij, L. S. (1978). *Mind in society*. Cambridge, MA: Harvard University Press.

Weinert, F. E. (2001). Vergleichende Leistungsmessung in Schulen – eine umstrittene Selbstverständlichkeit. In F. E. Weinert (Ed.), *Leistungsmessungen in Schulen* (pp. 17–31). Weinheim: Beltz.

Wiesner, C., Schreiner, C., & Breit, S. (2015). *Rahmenmodell zur pädagogischen Nutzung der Kompetenzorientierung durch die Bildungsstandardüberprüfung* (Unpublished paper). BIFIE, Salzburg.

Wiesner, C., Schreiner, C., Breit, S., & Bruneforth, M. (2017). Komplementäres Zusammenwirken von Standardüberprüfung und Informeller Kompetenzmessung. *Bifie-Journal*. https://www.bifie.at/wp-content/uploads/2018/03/bifie_journal_1-2017-04.pdf. Accessed May 17, 2018.

Open Access This chapter is licensed under the terms of the Creative Commons Attribution 4.0 International License (http://creativecommons.org/licenses/by/4.0/), which permits use, sharing, adaptation, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons licence and indicate if changes were made.

The images or other third party material in this chapter are included in the chapter's Creative Commons licence, unless indicated otherwise in a credit line to the material. If material is not included in the chapter's Creative Commons licence and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder.