

# Chapter 40

## Technology-Based Diagnostic Assessments for Identifying Early Mathematical Learning Difficulties



Gyöngyvér Molnár and Benő Csapó

### Introduction

The work presented in this chapter is located in the overlapping area of four research and development domains of education which have recently received growing attention. (1) Information communication technologies (ICT) have proliferated in all areas of life, including school learning. The ubiquitous ICT has made it more realistic to transfer all assessment to computerised platforms; therefore, *technology-based assessment* can be widely utilised to support everyday educational processes (Csapó, Ainley, Bennett, Latour, & Law, 2012). (2) Adapting education to the individual needs of students, and thus giving special support to those who really need it, has always been an intention of educators, but such a goal required assessment instruments that could diagnose students' difficulties early enough and monitor their progress. Therefore, assessment *for* learning, i.e. formative and *diagnostic assessment*, has recently become a dominant field within the research on educational assessment (Black & Wiliam, 1998), especially technology-based assessment. (3) Research has shown that preschool development and the first school years determine later success (Aunola, Leskinen, Lerkkanen, & Nurmi, 2004; Watts, Duncan, Siegler, & Davis-Kean, 2014); therefore *early childhood education* is one of the most rapidly growing areas in educational research, which development is strongly supported by the means of technology-based formative and diagnostic assessment. (4) Finally, mathematics is one of the most important school subjects;

---

G. Molnár (✉)

Institute of Education, University of Szeged, Szeged, Hungary

e-mail: [gymolnar@edpsy.u-szeged.hu](mailto:gymolnar@edpsy.u-szeged.hu)

B. Csapó

Institute of Education, University of Szeged, Szeged, Hungary

e-mail: [csapo@edpsy.u-szeged.hu](mailto:csapo@edpsy.u-szeged.hu)

success in learning it has a strong impact on a number of other areas of education, including science. Due to its importance in modern societies in everyday life as well as in science and technology-related professions, *mathematics education* has become one of the focal areas in improving educational systems. Besides reading and science, it is one of the three most frequently tested domains, both in international and national assessment programmes, thus attracting broad public attention.

There are many initiatives in progress, and a number of computer-based tests are available in the field of mathematics, but they are mainly developed for summative assessment, as well as the large-scale international (OECD PISA, IEA TIMSS) and national assessment programmes (e.g. MAP, Missouri Assessment Program, Missouri; SOL, Standards of Learning, Virginia; OAKS, Oregon Assessment of Knowledge and Skills, Oregon; SBAC, Smarter Balanced Assessment Consortium; PARCC, The Partnership for Assessment of Readiness for College and Careers). There are much fewer formative or diagnostic tests available, especially for measuring younger students' mathematics knowledge and skills. In general, there is a lack of research-based online diagnostic mathematics tests available for everyday classroom applications. Although there are several initiatives for online assessments (see, e.g. Pearson's MyMathLab (n.d.); Let's Go Learn (n.d.); The Diagnostic assessment part of PARCC (n.d.); Math Garden from the University of Amsterdam (n.d.); PAT: Mathematics in New Zealand (n.d.)), but these are all commercial products not completely and freely available for students and teachers.

As an exception, Panamath is available for free, but it is measuring only one part of mathematics' knowledge, students' approximate number system (ANS) aptitude. In the Panamath tasks, students are presented two sets of dots (blue and yellow), and they have to decide in a brief flash whether the number of blue or yellow dots is greater. The result tells about the accuracy of the test takers' basic sense for numbers. It can be used across the entire lifespan from 2-year-olds to old adults (DeWind & Brannon, 2016).

As there are large differences between students in a number of dimensions, successful mathematics education, especially in the first school years, requires differentiated and personalised teaching. This includes early identification of learning difficulties, frequent feedback, individualised well-targeted interventions, and continuous monitoring of development. An assessment system which can diagnose learning difficulties and can be used frequently enough must be built on a deeper understanding of students' developmental processes, the impacts of mathematics education on it, and the organisation of students' knowledge in general.

The first part of this chapter presents the advantages and possibilities of technology-based assessment. It describes how technology and its advantages initiated to rethink the purpose of assessment focusing more on diagnostic instead of summative assessment and realising efficient testing for personalised learning. The second part of the chapter summarises the scientific foundations for the diagnostic assessments. The theoretical foundations of framework development have resulted in a three-dimensional framework that outlines mathematics learning and the development of mathematical abilities and skills in three dimensions. These dimensions cover students' psychological development, the applicability of their

knowledge, and the curricular content of teaching. To implement the diagnostic assessment, a complex online platform called *eDia* has been constructed to support the entire assessment process from item writing through item banking, test delivery, and storing and analysing the data to providing feedback to students and their teachers. The third part of the chapter shows how the mathematics framework has been mapped into an item bank containing over a thousand items by dimension for the first six grades of primary school. The diagnostic assessment system has been offered to schools for application in everyday practice. The fourth part shows the implementation process, some early results from field testing, scaling issues, and framework validation. Finally, the last part discusses how the system can be further developed and how it can be integrated into everyday educational processes to support personalised education and provide customised support for atypical learners of mathematics.

### **Advantages and Possibilities of Technology-Based Assessment: The Move from Summative to Diagnostic Assessment to Realise Efficient Testing for Personalised Learning**

The most prominent educational developments of the past few decades have been aimed at establishing the feedback mechanisms of different levels of educational systems. Therefore, both the theory and the practice of educational assessment have seen considerable advances. Large-scale international assessments have become regularly administered by collaborative teams of experts of the leading test centres of the world. As a result, a huge improvement of data transfer technology and data analysis methods could be witnessed. Systems of assessment and evaluation in national contexts taking into account both the international trends and the local characteristics have been gradually set up. Due to the rapid development, the means of paper-based assessments most widespread and accepted at the millennium imposed serious constraints on their usability. To facilitate potential improvement and meet the twenty-first century needs of the new kinds of assessment and evaluation, an essential qualitative change had to be made (Scheuermann & Pereira, 2008). The direction of the change was mainly determined by technology. The fact that technology has developed, spread, and become accessible offers extraordinary opportunities for the improvement of the practice of educational assessment. Applying technology allows more exact and more varied testing procedures of significantly more complex skills and abilities by devising tasks in more realistic, application-oriented, and authentic testing environments than those of the earlier, paper-based assessments (Beller, 2013; Bennett, 2002; Breiter, Groß, & Stauke, 2013; Bridgeman, 2010; Christakoudis, Androulakis, & Zagouras, 2011; Csapó, Ainley, et al., 2012; Farcot & Latour, 2009; Kikis, 2010; Martin, 2010; Martin & Binkley, 2009; Moe, 2010; Ripley, 2010; van Lent, 2010). Its effectiveness and the increase of effectiveness under certain conditions could be detected on every level of assessment and evaluation.

- The economy of testing (Bennett, 2003; Choi & Tinkler, 2002; Farcot & Latour, 2008; Peak, 2005).
- The diversity of test editing and development (Csapó, Ainley, et al., 2012) and the speed of test administration and data flow (Csapó, Lőrincz, & Molnár, 2012).
- The opportunity to provide instant, objective, and standardised feedback (Becker, 2004; Dikli, 2006; Mitchell, Russel, Broomhead, & Aldridge, 2002; Valenti, Neri, & Cucchiarelli, 2003).
- The motivation of the students for testing changes (Meijer, 2010; Sim & Horton, 2005).
- Innovative item development opportunities, multimedia, dynamic, and interactive items, applying second- and third-generation tests (Pachler, Daly, Mor, & Mellar, 2010; Strain-Seymour, Way, & Dolan, 2009), which were impracticable in a paper-based form (Molnár, Greiff, Wüstenberg, & Fischer, 2017).
- An adaptive test algorithm has become available, which allows a more exact assessment of levels of knowledge and skills and abilities (Frey, 2007; Jodoin, Zenisky, & Hambleton, 2006).
- The circle of test takers could be extended (e.g. audio version of tasks and instructions could be played, which makes testing of children who cannot read possible) (Csapó, Molnár, & Nagy, 2014).
- Technology serves as an effective means of logging and analysing contextual data (e.g. the time needed for the execution of a task could be measured; besides the number of attempts made by the student to modify their solutions, the number and location of a student's clicks during a test could also be mapped) (Csapó et al., 2014). Consequently, instead of the only indicator used in paper-based testing, which is the test result, a rich and well-structured database is available, which makes a more thorough following and analysis of the student's movements and behaviour possible during the test (Molnár & Lőrincz, 2012).
- Indicators of test goodness criteria could increase (Csapó et al., 2014; Jurecka & Hartig, 2007; Ridgway & McCusker, 2003).

Although approaching the problem from different perspectives, major relevant research and development projects in an international context (e.g. *Assessment and Teaching of 21st Century Skills* – ATC21S, Class of 2020 Action Plan; Griffin, McGaw, & Care, 2012; SETDA, 2008) have all agreed that the direction for improvement could be computer-based testing exclusively (Csapó, Ainley, et al., 2012; Pearson, 2012; Scheuermann & Björnsson, 2009). Today computer-based assessment permits more effective assessments than traditional face-to-face or paper-based testing. Therefore, within a reasonably foreseeable time, all important assessment will probably be put on a technological basis. International summative tests have already shown such a tendency. Furthermore, given the opportunity to provide instant feedback on assessments, besides the predominantly summative approach, recently, there has been an emphasis on individualised diagnostic testing in order to enhance fast and effective learning by means of exploiting the learning supporting function of diagnostic testing (Kettler, 2011; Redecker & Johannessen, 2013; Van der Kleij, Eggen, Timmers, & Veldkamp, 2012). Traditional paper-based

tests are not suitable for diagnostic assessment, which bottom line is sufficiently frequent student assessment. The development of technology together with that of assessment and evaluation in the past 15 years has created numerous new opportunities in early childhood assessment, which so far have mainly been based on individual data collection (Csapó et al., 2014).

## **Theoretical Foundations of Framework Development: A Three-Dimensional Model of Mathematical Knowledge**

In the history of mathematics education, three perennial goals have remained clear from the very beginning of the history of schooling up to present-day approaches. To create a diagnostic assessment system which can precisely identify students' weaknesses and strengths, a framework must be created which clearly distinguishes these three directions, three types of goals.

Cultivating general cognitive abilities has always been one of the main declared goals of learning mathematics. Adjusting learning to students' mental development is a precondition of successful teaching, while obtaining feedback on how maths teaching stimulates the developing mind requires regular testing. To create assessment instruments to meet this goal, psychological processes must be studied.

Another obvious goal is that mathematics education should provide learners with practical skills applicable outside the school context. Seneca's often cited aphorism, "Non scholae sed vitae discimus", expresses the expectations of modern societies as well, and this aim, making mathematics education more relevant for the average learner, is embodied in national and international assessment projects.

Finally, mathematics is one of the oldest and best organised bodies of human knowledge. As Banach has formulated his admiration, "Mathematics is the most beautiful and most powerful creation of the human spirit". To comprehend the organisation of this branch of knowledge, students must study mathematics as a discipline, including its specific terminology, its axioms, theorems, definitions, proofs, etc. Another set of goals can be deduced from this need which can be further shaped taking into account the educational requirements of those students who prepare to be professional users of mathematics, becoming research mathematicians or dealing with high-level applications in a number of other areas of research and development.

### ***A Three-Dimensional Model of Students' Knowledge for Diagnostic Assessment in Early Education***

The arguments for assessing students' progress in three dimensions may be further elaborated by analysing some national and international assessment frameworks. Large-scale international assessment programmes publish their frameworks well

before the actual assessment. The first international assessment programmes in mathematics were conducted by the IEA (International Association for the Assessment of Educational Progress) in the early 1970s and 1980s, and the assessments became regular since 1995 under the acronym TIMSS (Trends in International Mathematics and Science Studies). The early IEA assessments focused on the curricular content of mathematics teaching and were closer to the disciplinary view of mathematics. Although the curricula in the participating countries remained the primary source of content for recent TIMSS assessments, they distinguish the content domains (covering the main domains of mathematics as a discipline) and the cognitive domains which are knowing, applying, and reasoning (Mullis & Martin, 2013).

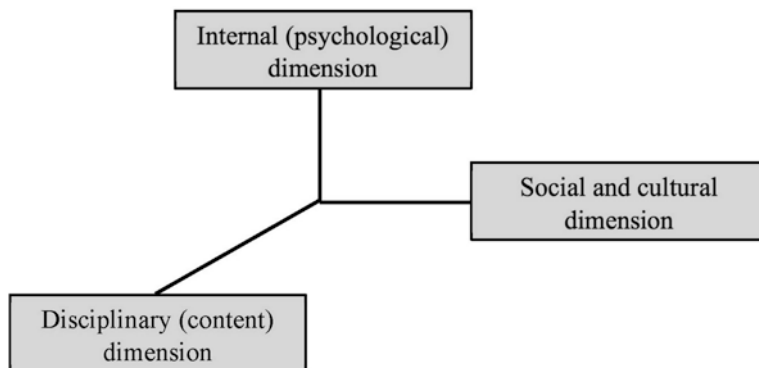
The other large-scale international programme launched in 2000 under the aegis of the OECD, PISA (Programme for International Student Assessment), aims to assess the knowledge and skills that students are expected to possess at the age of 15 to be prepared for the challenges they will face in modern societies. To characterise the type of broadly applicable knowledge, PISA extended the conception of *literacy* and termed the assessment domains reading literacy, mathematical literacy, and scientific literacy. For mathematical literacy, a novel definition was developed:

Mathematical literacy is defined in PISA as: the capacity to identify, to understand, and to engage in mathematics and make well-founded judgments about the role that mathematics plays, as needed for an individual's current and future private life, occupational life, social life with peers and relatives, and life as a constructive, concerned, and reflective citizen. (OECD, 2000, p. 50)

Based on this definition, the framework was elaborated in three dimensions, dealing with *mathematical processes*, *mathematical content*, and *situations and contexts* of applying mathematical knowledge. Both the definition of mathematical literacy and the detailed framework proceeding from it as well as item development placed much stronger emphasis on the application of knowledge as the disciplinary content and the mathematical processes were embedded in contexts and situations relevant for young students living in developed societies. Over the assessment cycles, the conception of mathematical literacy further evolved, and its core idea remained very similar to the original:

Mathematical literacy is an individual's capacity to formulate, employ and interpret mathematics in a variety of contexts. It includes reasoning mathematically and using mathematical concepts, procedures, facts and tools to describe, explain and predict phenomena. It assists individuals to recognize the role that mathematics plays in the world and to make the well-founded judgments and decisions needed by constructive, engaged and reflective citizens. (OECD, 2016, p. 65)

Following the traditions of framework development in international assessment projects and taking into account several further theoretical considerations (see Csapó, 2004, 2010) and empirical results (Csapó, 2007), a three-dimensional model of teaching and learning goals was proposed. This approach (outlined in Fig. 40.1) assumes that these three aspects of teaching should be present at the same time in school education, to develop the intellect and to cultivate thinking and general cognitive abilities. These goals must not exclude each other, and they should not compete for teaching time. Focusing on one of these goals, e.g. teaching the disciplinary



**Fig. 40.1** A three-dimensional model for developing a framework of diagnostic assessments. (Csapó, 2007)

content (which traditionally happens in many education systems), is not satisfying in modern societies; students are expected to apply their mathematical knowledge in a broad variety of contexts (as PISA assesses it), and they should be able to solve problems in unknown, novel situations (as was assessed, e.g. in PISA 2012 in the domain of problem-solving, see OECD, 2014). These goals (teaching disciplinary content knowledge in mathematics, preparing students to apply it in a broad range of contexts, and developing thinking skills; see Csapó & Szendrei, 2011) have been competing with each other for teaching time over the past few decades. One or another became from time to time dominant in the curricula; however, they should receive equal attention for interacting and reinforcing each other.

The model in Fig. 40.1 has been elaborated for each assessment domain taking into account the specific characteristics of the particular domain and has been published in three parallel volumes (see Csapó & Csépe, 2012 for reading; Csapó & Szendrei, 2011 for mathematics; and Csapó & Szabó, 2012 for science). The similarities and differences of these frameworks highlighted the specific roles each domain plays in education. Reading is the basis for all further learning, including mathematics, while mathematics provides foundations for learning certain sciences. Further developmental work (creating items and carrying out assessments with them) based on this three-dimensional framework indicates the validity of the approach in educational practice.

As for mathematics, each dimension has been separately considered and elaborated in detail in the light of literature from the particular field of research. It is of great use to separate these different dimensions in diagnostic assessments because a precise identification of areas of delayed differences is a precondition of personalised interventions. The scope of studying these dimensions is also different. The roots of the psychological development of mathematical reasoning may be universal as far as early neurocognitive development in children is alike across cultures and societies. Studies related to the application dimension can mostly be shared with researchers dealing with the contexts and expectations of developed countries, while the curricular content is related to the national educational system.



The *psychological dimension* has been conceptualised as the interaction between students' cognitive development and learning mathematics at school (Nunes & Csapó, 2011). The questions in this dimension are how well mathematics education is adjusted to students' psychological development, on the one hand, and how learning mathematics can contribute to the development of specific reasoning skills and how effectively it stimulates students' general cognitive development, on the other. Research in this field provides rich resources ranging from the classical works of Piaget (see, e.g. Inhelder & Piaget, 1958) to the most recent neurocognitive studies. A long list of skills can be taken into account in this field that are strongly embedded in psychological development, such as counting skills, additive and multiplicative reasoning as well as spatial, probabilistic, combinatorial, and proportional reasoning, and so on. Assessments of a number of such skills are especially crucial at the beginning of schooling and in the first school years, as their developmental level determines later success (see Nguyen et al., 2016).

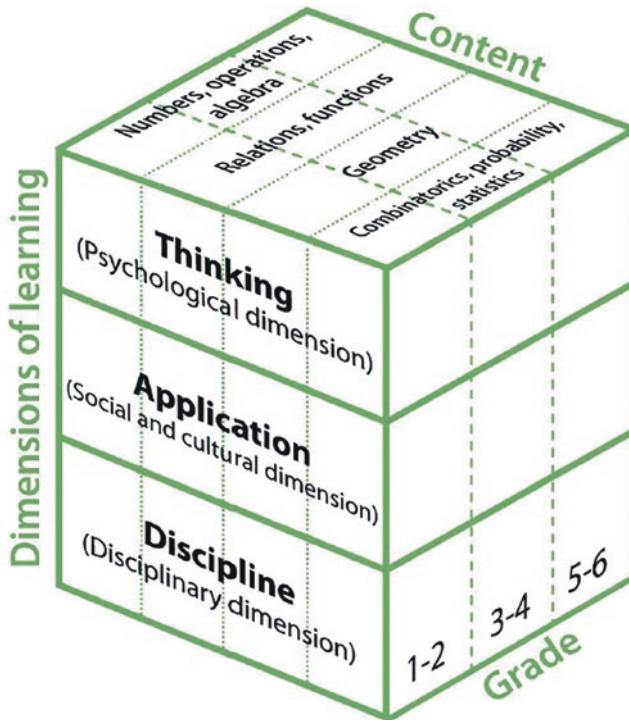
The *application dimension* of the goals of learning mathematics is interpreted as mastering mathematical literacy, the type of skills that make mathematics useful in areas other than the immediate school context. Mathematics is applied in a number of areas, ranging from other school subjects to a broad cross section of everyday life (Csíkos & Verschaffel, 2011). The key questions in this field are how students can construct mathematical models of problems they face and how well they can mobilise mathematical knowledge to solve those problems. Transfer of knowledge to new contexts is not automatic, and children must learn and practise applying their knowledge. Research on realistic mathematical modelling is the most useful source for elaborating the assessment framework of this dimension (see, e.g. Verschaffel, De Corte, & Lasure, 1994). The tasks that can be taken into account for the measurement of this dimension range from pseudo-real-world to real-world problems which embed mathematical knowledge in a number of relevant contexts and real-life situations.

The *disciplinary dimension* can be defined as the mathematics content knowledge described in the national core curriculum. This is the prescribed content on which textbooks, local curricula, and teachers' actual work are based (Szendrei & Szendrei, 2011). A precise translation of the core curriculum into an assessment framework and later on into test tasks makes it possible to monitor how students progress with their daily mathematics studies. Previous research has indicated that mastering and reproducing the immediate teaching material does not necessarily have a long-term impact on students' cognitive development (see, e.g. Csapó, 2007), but for a precise diagnosis, it is necessary to know if students actually learn what they are expected to in mathematics lessons.

Teaching students disciplinary content knowledge in mathematics, preparing them to apply it, and developing their thinking skills are not considered as exclusive alternatives but processes that reinforce and interact with each other. That is, education must achieve these objectives in an integrated way, but for diagnostic purposes the tests must be able to show if there is insufficient progress in one or another of these dimensions, thus they should be treated as distinct dimensions in diagnostic assessments.

Taking this principle into account and considering the specific aspects of early education and the diagnostic orientation of the assessment, the former tridimensional





**Fig. 40.2** The model of mathematical knowledge to develop the framework for diagnostic assessment in Grades 1–6. (Csíkos & Csapó, 2011)

model was further developed and used as a foundation for item development. The continuum of the first six grades has been divided into 2-year sections, and the test items have been prepared to cover these periods (Csíkos & Csapó, 2011, see Fig. 40.2).

The three-dimensional approach indicates that these aspects of learning are not independent of each other. Disciplinary content is the means of developing students' reasoning skills, and this is what students are then expected to apply in other contexts. The following sections show how items were developed for these dimensions, how students' knowledge is measured in these dimensions, and how their disciplinary knowledge, reasoning skills, and applicable knowledge are related.

## Creating an Assessment System: Online Platform Building and Innovative Item Writing

Based on the model of mathematical knowledge described in the previous section, an item bank was constructed for diagnostic assessments. This item bank contains 6182 tasks (each task consists of several items) to measure disciplinary content

knowledge in mathematics (MD;  $n = 2119$ ), mathematical reasoning (MR;  $n = 1965$ ), and mathematical literacy (MA;  $n = 2098$ ) in first to sixth grades (age 6–12). The content of the assessment as a function of the three dimensions of learning and target population is shown in Fig. 40.2.

The tasks were grouped into clusters (4–5 tasks per cluster), meaning 15–20 items per cluster for the lower grades and 20–25 items for the higher grades. One 45-min test contains at least three clusters (at least 45–50 items).

In the first to third grades, instructions are provided both in written form and online by a prerecorded voice to prevent reading difficulties and ensure the validity of the results. Thus, students must use headphones during the administration of the tests. After listening to the instructions, they must indicate their answer using the mouse or keyboard (in the case of desktop computers, which is the most common infrastructure in the Hungarian educational system) or directly tapping, typing, or dragging the elements of the tasks with their fingers on tablets. It takes no more than 45 min (one school lesson) to complete the test.

At the beginning of the tests, participants are provided with instructions, including a trial (warm-up) task with immediate feedback, in which they can learn how to use the programme: (1) at the top of the screen, a yellow bar indicates how far they have advanced in the test; (2) they must click on the speaker to be able to listen to the task instruction; (3) to move on to the next task, they must click on the “next” button; and, finally, (4) after completing the last task, they receive game-based immediate feedback with one to ten balloons depending on their achievement. The better their results are, the more balloons they will see over Piglet’s head. The immediate feedback also contains their achievement in each dimension of knowledge.

The feedback system, which is available for the teacher, is more elaborated. As the tasks in the item bank have been scaled by means of IRT, students’ achievement can be objectively compared. Teachers receive feedback on students’ achievements both in percentage and in ability scores, which are comparable to each other and also contain a point of reference to the national standards. In each of the grades and fields, the national-level average achievement was transferred to 500 points with a 100-point standard deviation, which constructs the point of reference to the students’ achievement.

### ***Mathematical Reasoning Items***

Based on the framework for the diagnostic assessment of mathematics (Csapó & Szendrei, 2011), reasoning items encompass the measurement of inductive reasoning, deductive reasoning, combinative reasoning, systematisation skills, and proportional reasoning. The task presented in Fig. 40.3 combines the mathematical concept of whole numbers with the assessment of students’ inductive reasoning skills within the context of a famous Hungarian cartoon. In the task, students must

Dumpling Arthur got 20 bars of chocolate for his birthday. He ate a few pieces from each chocolate bar, and then he put the rest of the chocolate into groups according to a certain rule. There is an odd one out in each row. Which one is it? Click on it.


←

→

**Fig. 40.3** Mathematical reasoning task: combining the mathematical concept of whole numbers with the assessment of students' inductive reasoning skills in a familiar Hungarian cartoon

discover regularities by detecting dissimilarities with respect to attributes of different objects. In this operation, they must use their knowledge of quantities and their understanding of the relations of greater than, less than and the same. According to Klauer's definition (1993) of inductive reasoning, students must use the operation of discrimination in this item. In the present case, students can provide their answers by clicking on the "odd-one-out" element, scoring a maximum of 4 points, one in each group of chocolate bars. As demonstrated, inductive reasoning tasks are often connected to other areas of mathematics, in this case to whole numbers and computation.

In the task presented in Fig. 40.4, students' systematisation skills and their level of understanding of the number concept are assessed. The formulation and development of the number concept must be supported from three directions: number symbols, the name of the numbers, and the quantities indicated by the numbers.

The tasks, which support the connections between these representations, are suitable for diagnostic purposes for the reasoning dimension of counting. The present task provides an example of the combination of number symbols and quantities in a reasoning context. Students need to recognise number symbols and then connect them to quantities and place them in increasing order by clicking on the numbers. In short, the order of the clicking was evaluated.

Rose Frog is jumping from one lily pad to the next with the numbers on the lily pads going up in each row.

Click on the lily pads with the numbers on them going up in each row.

Row 1: 4, 7, 13, 10

Row 2: 7, 12, 5, 16

Row 3: 9, 18, 8, 19


Row 4: 10, 20, 4, 14


**Fig. 40.4** Mathematical reasoning task: recognising and combining number symbols and quantities

### *Mathematical Literacy Items*

In the lower grades, mathematical problems become realistic when everyday experiences and observations come to play an active role in the problem-solving process. It is easier to interpret the problem if it is supported by a relevant picture or situation. The word problems can be made realistic if they can be solved with the accompanying picture or by manipulating the pictures given. The task presented in Fig. 40.5 using online technology encompasses an important feature of an authentic problem beyond the real-life-like context; namely, several solutions are possible, and the students can interact with the problem environment. With the scoring procedure, it is all the same which of the teddy bears are placed – dragged and dropped – on the bed; only the number of teddy bears counts. All of the combinations are accepted. The task measures skill level addition up to 10 in a realistic application context.

The task presented in Fig. 40.6 illustrates that it is impossible to split the tasks of the three dimensions from each other. It is a mathematical literacy type task, which measures number concept and relations and functions in a realistic application context. The aim is to measure students' ability to follow, recognise, and continue periodically repeating rhythms and movements by detecting similarities in relations among objects in an application context. Task scoring is automatised for all of the

 Keep putting Teddy bears on the bed till you have 8 bears there.



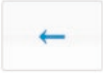
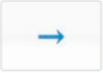



 

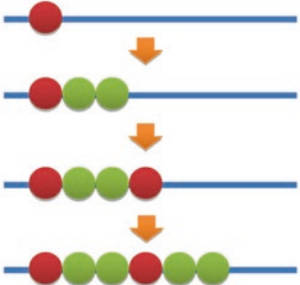

Fig. 40.5 Mathematical literacy task: adding up to ten in a realistic application context

 Ann is making a bracelet out of beads for her friend following the pattern in the drawing. When she is ready, the bracelet will have 12 beads.

 How many red beads did she use? Write the correct number in the text box.

She needs  red beads for the bracelet.

 What is the colour of the last bead Ann used? Click on the picture.



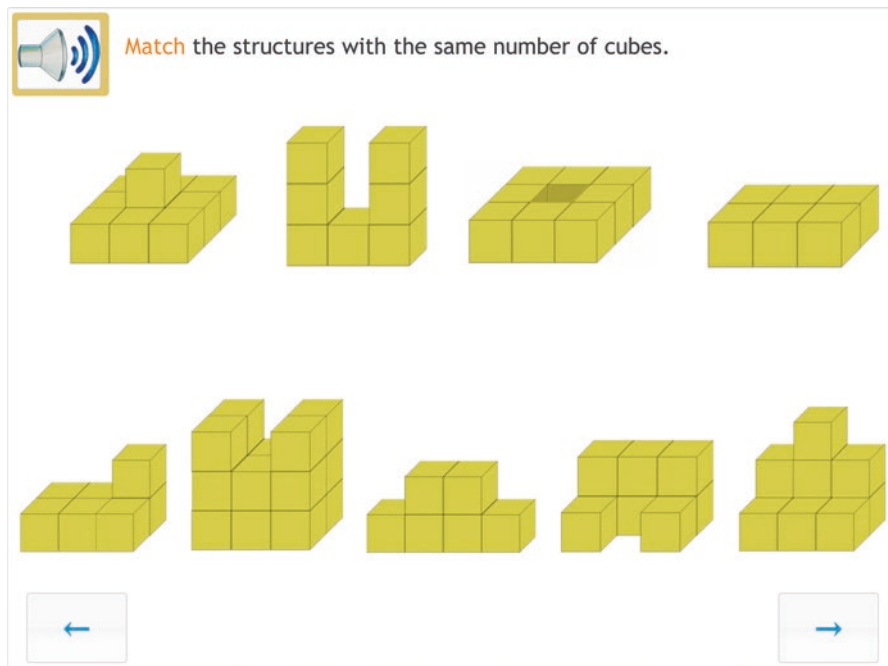
 

Fig. 40.6 Mathematical literacy task: following, recognising, and continuing periodically repeating rhythms and movements in a realistic application context

tasks in the item bank, even those with several correct answers. In the present case (Fig. 40.6), students had the option of typing their answer for the number of beads used in the bracelet in several ways, e.g. using number symbols or letters, using small letters or capitals or a mixture of them, or using spaces. All of these possibilities were accepted by the scoring system.

### Items that Assess Disciplinary Mathematics Knowledge

In early mathematics education, among the most effective teaching methods are learning-by-doing activities. This is also the case for geometry, where students need to discover three-dimensional forms through different activities. The experience gained during these activities provides the foundations and in many cases determines the conceptual building work in lower and higher grades. In an online environment, the possibilities of manipulation play an important role. The task presented in Fig. 40.7 illustrates this. Students need to connect three-dimensional forms built out of cubes with other three-dimensional forms consisting of the same number of cubes by clicking on them to draw the connections. As GeoGebra elements and tasks uploaded from GeoGebraTube can be used in the *eDia* system, students can even rotate and engage in a manipulative interaction with these three-dimensional geometric forms.



**Fig. 40.7** Mathematical disciplinary task from geometry involving learning-by-doing with the possibility of manipulative interaction



The different colours tell you who made the throw.

Click on the name of the person with the least points.

Who got more points?

True or false? Mark the sentences T (true) or F (false).

John has one more point than Mark. Click here!

Mark has three more points than Paul. Click here!

Legend:  
 John (red dot)  
 Mark (brown dot)  
 Lisa (blue dot)  
 Mary (yellow dot)  
 Paul (green dot)

Target Diagram Data:  
 Ring 1 (outermost): 1 point  
 Ring 2: 2 points  
 Ring 3: 3 points  
 Ring 4: 4 points  
 Ring 5 (innermost): 5 points

Points per child:  
 John: 1 point (Ring 1)  
 Mark: 1 point (Ring 1)  
 Lisa: 1 point (Ring 1)  
 Mary: 1 point (Ring 1)  
 Paul: 1 point (Ring 1)

**Fig. 40.8** Mathematical disciplinary task: integrating the understanding of number symbols, the operation of addition, the comparison of quantities and numerosities, and the knowledge of relation symbols

In the first few years of schooling, operations with whole numbers, which build the foundations for additive reasoning, form an essential part of mathematics education. They include not only the operation of addition but all the knowledge elements for comparing quantities and numerosities. By reading the different numbers, sums, and differences, students are prepared for the mathematical concepts of addition and subtraction. In the process of understanding and interpreting addition and subtraction, the number line plays an important role. The task presented in Fig. 40.8 integrates the understanding of number symbols, the operation of addition, the comparison of quantities and numerosities, and the knowledge of relation symbols. During the solution process, students had to click on the name of the child who scored the least points and then choose the right relation symbol from the drop-down menus.

### Field Trial and Empirical Validation of the Theoretical Model

We launched a field trial study to ascertain the applicability of computer-based tests in regular educational practice for assessing students at the beginning of schooling and for the empirical validation of the theoretical model of mathematical knowledge



introduced above. The objectives of the study were threefold. First, we examined the applicability of an online diagnostic assessment system in the field of mathematics for students at the beginning of schooling. We then empirically validated the three-dimensional model of mathematical knowledge based on research results collected with first graders using *eDia*, the Hungarian online diagnostic assessment system. Finally, we examined the relationship between disciplinary content knowledge (MD), mathematical reasoning (MR), and mathematical literacy (ML) and answered the research question: how are the three different dimensions of mathematical knowledge related?

The sample was drawn from first-grade students in Hungarian primary schools. School classes formed the sampling units. 5115 first graders were involved in the study. The proportion of girls and boys was about the same.

The instrument was only a part of the whole test battery; it consisted of 48 items, which measured MD, ML, and MR in that order. To prevent reading difficulties, instructions were provided online using a prerecorded voice. Children had to indicate their answer by using the mouse or keyboard. Testing took place in the computer labs at the participating schools. Test completion lasted no more than 45 min (one school lesson). The tests were automatically scored, and students received immediate feedback at the end of the testing.

Reliability, time-on-task, and missing and achievement data were analysed to test the applicability of the online assessment system by first graders. The Rasch model was used to scale the data and draw the three-dimensional item-person map of mathematics. We conducted confirmatory factor analyses (CFA) within structural equation modelling (SEM; Bollen, 1989) to test the underlying measurement model of mathematical knowledge with the three different dimensions: disciplinary knowledge, literacy, and reasoning. Bivariate correlations, partial correlations, and SEM analyses were employed to test construct validity, that is, the relations between the three dimensions of mathematical knowledge.

Why have we conducted confirmatory factor analyses and what is it good for? Confirmatory factor analysis is a special form of factor analyses. In the present case, it is used to test whether the model based on the empirical data is consistent with our understanding of the nature and of the three-dimensional model of mathematical knowledge. That is, the objective of confirmatory factor analysis is to test whether the data fit a hypothesised measurement model, which is based on the three-dimensional theory of knowledge.

Bivariate correlation indicates the numerical relationship, the strength of the association between two measured variables, while partial correlation measures the degree of this association with the effect of controlling variables removed. Bivariate correlations can give misleading results if there is another variable that is related to both of the examined variables. This misleading information can be avoided by computing the partial correlation coefficient. Both of the coefficients take on a value in the range from  $-1$  to  $1$ . The value  $0$  conveys that there is no relationship, the value  $-1$  means a perfect negative correlation, and the value  $1$  conveys a perfect positive association.

Construct validity describes the degree to which a test measures what it claims, indicating how well it really covers the targeted content; whether the scale behaves like the theory predicts a measure of that construct should behave. It describes the degree to which empirical evidence and theoretical rationales support the adequacy and appropriateness of inferences and actions based on test scores (Messick, 1995).

### ***Applicability of the Diagnostic System in Everyday School Practice***

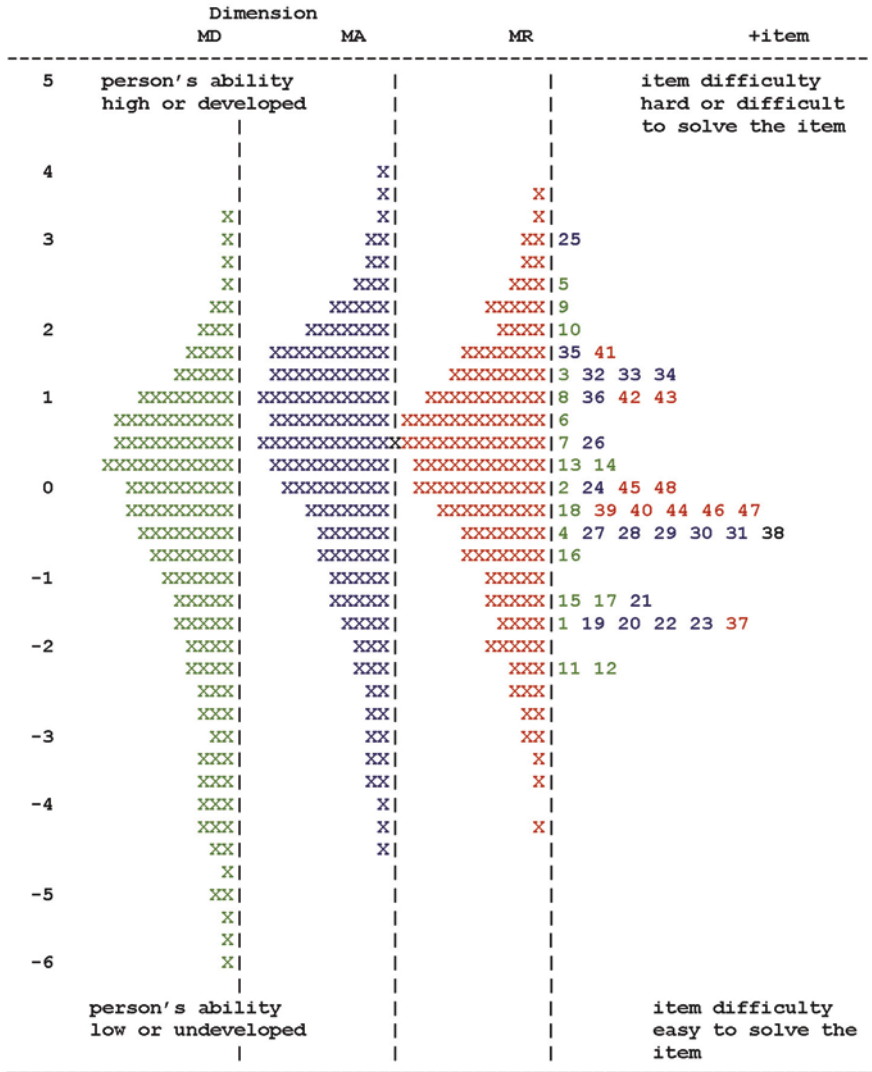
The results confirmed our hypotheses. The internal consistency of the mathematics test proved to be high both on the test ( $\alpha = 0.942$ ) and subtest levels ( $\alpha_{MD} = 0.89$ ;  $\alpha_{MR} = 0.83$ ;  $\alpha_{ML} = 0.89$ ), so the results are reliable and generalisable. Less than 0.4%, that is, 18 students out of 5115, were not able to finish the test on time (within 45 min). As none of them completed more than 70% of the test and reached the third subtest, all of their data were deleted from the databases that form the data for the 5097 students involved in the analyses. Generally, the students managed to finish the test within the given timeframe, 1690 seconds on average ( $sd = 673$ ).

### ***Scaling and Item Difficulty***

Participants' score distribution on the mathematics test also confirmed the applicability of the online assessment system. The mean achievement was about 50% (49.39%,  $sd = 23.87$ ). The subtest level achievement distribution changed ( $M_{md} = 42.29$ ,  $s = 26.66$ ;  $M_{ml} = 53.96$ ,  $s = 26.67$ ;  $M_{mr} = 53.18$ ,  $s = 28.14$ ) and was significantly different ( $t_{md\_ml} = -40.96$ ,  $p < 0.01$ ;  $M_{md\_mr} = -33.45$ ,  $p < 0.01$ ;  $M_{ml\_mr} = 2.35$ ,  $p < 0.05$ ). The level of standard deviations indicated that the test could be used to test the variability of the sample even on a subtest level.

The three-dimensional item-person map (Fig. 40.9) shows the match between the item difficulty distribution and the distribution of students' Rasch-scaled achievement estimates for MD, MA, and MR. For any person engaged with an item located at that person's level, the Rasch model routinely sets the probability of success on the item at 50% on an item-person logit scale.

The probability of success increases to 75% for an item that is 1 logit easier or decreases to 25% for an item that is 1 logit more difficult. The MD (green signs) and MA (blue signs) items were well matched to the sample ("x" and number are parallel), and with MR some hard and easy items were missing from the test. The achievement distribution in MD was the highest; there were more low-developed students than in the two other dimensions. Generally, the test was suitable for measuring and discriminating student achievement based on the three-dimensional model of mathematical knowledge in first grade in an online environment.



Each 'X' represents 35.0 cases

Fig. 40.9 The three-dimensional item-person map of first graders' mathematical knowledge

Gender-level achievement differences changed between the different dimensions. Girls' achievement proved to be significantly higher on the test level ( $M_{girl} = 50.36$ ,  $s_{girl} = 23.46$ ,  $M_{boy} = 49.01$ ,  $s_{boy} = 23.92$ ,  $t = -2.011$ ,  $p = 0.044$ ); however, the level of significance might only have been caused by the large sample size. On the subtest level, there were no gender-level achievement differences on the MD and MA subtests, while significant differences could be detected on the reasoning part of the test ( $t = -2.923$ ,  $p < 0.01$ ), thus causing the gender-level differences on the test level.

Based on this result, we can conclude that, first, computer-based assessment can be carried out even at the very beginning of schooling without any modern touch screen technology on normal desktop computers using a general browser and the school infrastructure, and, second, the online diagnostic system can be used to test students' mathematics knowledge at the beginning of schooling in a school context.

### *Dimensionality and Structural Validity*

In validating the three-dimensional model of mathematical knowledge, SEM analyses were outperformed. The three-dimensional measurement model for mathematics showed a good model fit (Table 40.1), based on Hu and Bentler's (1999) recommended cut-off values. The comparative fit index (CFI) and the Tucker-Lewis index (TLI) value above 0.95 and the root mean square error of approximation (RMSEA) below 0.06 indicate a good global model fit. As significant and high correlations were found between the pairs of dimensions ( $r_{MD\_MR} = 0.685$ ,  $r_{MD\_ML} = 0.749$ ,  $r_{ML\_MR} = 0.634$ ,  $p < 0.001$ ) on a latent level – latent variables are not directly observed but are inferred from other variables that are observed (directly measured) – within the three-dimensional model, we also tested the one-dimensional model with the three dimensions combined under one general factor. With the one-dimensional model, the fit indices decreased considerably.

In order to test which model fitted the data better, a special  $\chi^2$ -difference test was carried out in Mplus, which showed that the three-dimensional model fitted significantly better than the one-dimensional model ( $\chi^2 = 3389.111$ ;  $df = 6$ ;  $p < 0.001$ ). In summary, the three-dimensional model fitted well and better than the one-dimensional model. Thus, the disciplinary, literacy, and reasoning dimensions of mathematical knowledge were empirically distinguished, supporting our hypothesis.

The bivariate correlations between MD, ML, and MR were high, ranging from 0.63 to 0.71 (Fig. 40.10). The relationships proved to be similar between MR and either ML or MD ( $r = 0.63$  and  $0.64$ ,  $p < 0.001$ , respectively), and they were significantly weaker than the correlation between ML and MD ( $r = 0.71$ ,  $p < 0.001$ ).

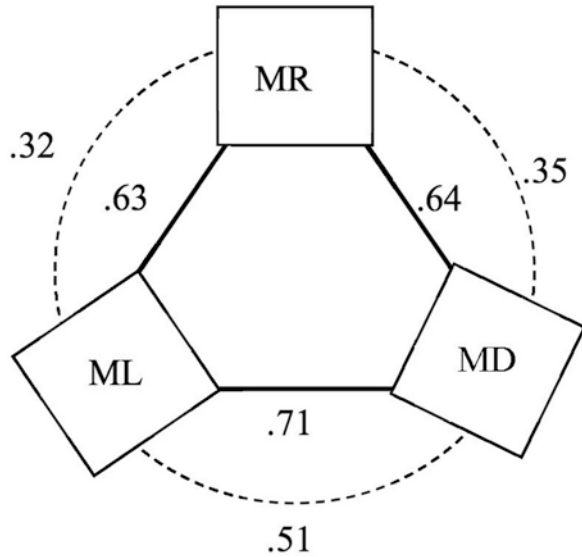
Partial correlations were significantly lower as all bivariate relationships were influenced by the third construct ( $r_{MR\_ML} = 0.32$ ;  $r_{MR\_MD} = 0.35$ ;  $r_{ML\_MD} = 0.51$ ,  $p < 0.001$ ). Like the bivariate correlations, the partial correlation coefficients between MR and either ML or MD were of the same strength ( $p < 0.001$ ), while the partial

**Table 40.1** Goodness of fit indices for testing dimensionality of mathematics

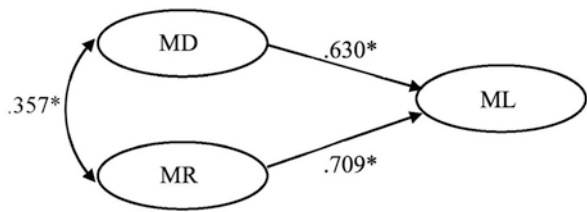
Model	$\chi^2$	<i>Df</i>	<i>p</i>	CFI	TLI	RMSEA (90% CI)	<i>n</i>
Three-dimensional	16955.213	1067	0.001	0.965	0.963	0.054 (0.053–0.055)	5097
One-dimensional	31445.929	1073	0.001	0.931	0.928	0.075 (0.075–0.076)	5097

Note: *df* degrees of freedom, *CFI* comparative fit index, *TLI* Tucker-Lewis index, *RMSEA* root mean square error of approximation,  $\chi^2$  and *df* are estimated by WLSMV

**Fig. 40.10** Relations between MR, ML, and MD (Solid lines depict bivariate correlations; dotted lines represent partial correlations. All coefficients are significant at the  $p < 0.001$  level)



**Fig. 40.11** A structural model of mathematical knowledge: disciplinary knowledge and mathematical reasoning as predictors of mathematical literacy ( $*p < 0.01$ )



correlation between ML and MD proved to be the highest. This is supported and was indicated by the correlation coefficients on a latent level as well (see above).

We assumed that disciplinary knowledge and mathematical reasoning predict performance in literacy, the application dimension of mathematics, since we need that dimension of mathematics most in everyday life. Thus, we regressed MD and MR on ML and estimated the proportion of variance explained. The results showed that MD and MR explained performance in ML on a high level (90%) but with a different effect (see Fig. 40.11). The residuals of measures of MD and MR were still correlated on a moderate level ( $r = 0.35$ ), indicating common aspects of MD and MR that are separable from ML. The model fit well (CFI = 1.000, TLI = 1.000, RMSEA = 0.000).

To sum up, our results showed that MD, MR, and ML are highly correlated constructs, though not identical. Students' levels of disciplinary knowledge and mathematical reasoning strongly influence and predict achievement in the context of mathematical application. That is, if we enhance disciplinary knowledge in mathematics and students' thinking skills, we can expect a stronger transfer from the disciplinary to the application contexts. This suggests that beyond factual

knowledge, thinking skills should become an integral part of school agendas (de Koning, 2000) and should be incorporated into a broad range of school-related mathematical learning activities.

## Conclusions and Further Research and Development

In this chapter, we have presented the theoretical foundations and technological realisation of an online diagnostic assessment system in the domain of mathematics. The applicability of this system in educational practice was demonstrated in an ecologically valid context, when the online tests were administered to a large sample in real school settings. The assumption that computer-based assessment is applicable even in the early school grades was confirmed. We validated the three-dimensional model of mathematical knowledge empirically, having addressed the psychological, application, and disciplinary dimensions of knowledge. These results strengthen the foundations for a complex online diagnostic assessment platform called *eDia*, which contains about 2000 tasks (8000 items) per dimension for the first six grades of primary school.

According to the empirical results, the three-dimensional approach is valid; the disciplinary, application, and reasoning aspects of learning are neither independent of nor identical to each other. Consequently, each of these three aspects of knowledge must be enhanced at the same level and at the same time at school, and all of them must be incorporated into a broad range of mathematical learning activities and must not be mutually exclusive. In modern societies, it is neither sufficient nor satisfying to focus on only one of these goals, a common tendency in many education systems in which the teaching of disciplinary content is favoured.

The system can be used to identify students with atypical development, that is, children whose achievement is significantly lower in one of the three dimensions. Teachers receive prompt feedback about their students' development in each of the dimensions separately in a comparable way. At this moment, in the phase of system development, the system administers the tests having different difficulty levels to the students in a random way; it is not enough to provide only percentage-based feedback to the teachers, as they are strictly taken not objectively comparable to each other. The feedback is based on students IRT-based ability levels in ability points, which can be referred to the national mean ability values that is transferred to 500 (with 100-point standard deviation) in each grade, which constructs the point of reference to the students' achievement. Beyond the student-level results and national standards, teachers receive feedback about their class-level and school-level achievement with comparison to the other class-level, school-level, regional-level, and strata-level achievements. Our future plan is to put the test administration on an adaptive level.

Training programmes adjusted to their specific deficiencies can then be implemented to help them catch up. The efficacy of such a training can also be monitored with the assessment system. Further research can be carried out with the diagnostic

assessment system to explore the reason for an atypical mathematical development and the ways in which the different dimensions of mathematical knowledge can be effectively enhanced.

**Acknowledgement** This study was funded by OTKA K115497.

## References

- Aunola, K., Leskinen, E., Lerkanen, M. K., & Nurmi, J. E. (2004). Developmental dynamics of math performance from preschool to grade 2. *Journal of Educational Psychology*, 96(4), 699–713. <https://doi.org/10.1037/0022-0663.96.4.699>
- Becker, J. (2004). Computergestütztes Adaptives Testen (CAT) von Angst entwickelt auf der Grundlage der Item Response Theorie (IRT). Unpublished PhD dissertation. Freie Universität, Berlin.
- Beller, M. (2013). Technologies in large-scale assessments: New directions, challenges, and opportunities. In M. von Davier, E. Gonzalez, I. Kirsch, & K. Yamamoto (Eds.), *The role of international large-scale assessments: Perspectives from technology, economy, and educational research* (pp. 25–45). Dordrecht, The Netherlands: Springer.
- Bennett, R. E. (2002). *Using electronic assessment to measure student performance. The state education standard*. Washington, DC: National State Boards of Education.
- Bennett, R. E. (2003). *Online assessment and the comparability of score meaning*. Princeton, NJ: Educational Testing Service.
- Black, P., & Wiliam, D. (1998). Assessment and classroom learning. *Assessment in Education: Principles, Policy & Practice*, 5(1), 7–74. <https://doi.org/10.1080/0969595980050102>
- Bollen, K. A. (1989). *Structural equations with latent variables*. New York, NY: Wiley.
- Breiter, A., Groß, L. M., & Stauke, E. (2013). Computer-based large-scale assessments in Germany. In D. Passey, A. Breiter, & A. Visscher (Eds.), *Next generation of information technology in educational management* (pp. 41–54). Berlin, Heidelberg: Springer.
- Bridgeman, B. (2010). Experiences from large-scale computer-based testing in the USA. In F. Scheuermann & J. Björnsson (Eds.), *The transition to Computer-based assessment: New approaches to skills assessment and implications for large-scale testing* (pp. 39–44). Brussels: European Communities.
- Choi, S. W., & Tinkler, T. (2002). *Evaluating comparability of paper and computer based assessment in a K-12 setting*. Paper presented at the annual meeting of the National Council on Measurement in Education, New Orleans, LA.
- Christakoudis, C., Androulakis, G. S., & Zagouras, C. (2011). Prepare items for large scale computer based assessment: Case study for teachers' certification on basic computer skills. *Procedia-Social and Behavioral Sciences*, 29, 1189–1198.
- Csapó, B. (2004). Knowledge and competencies. In J. Letschert (Ed.), *The integrated person. How curriculum development relates to new competencies* (pp. 35–49). Enschede: CIDREE.
- Csapó, B. (2007). Research into learning to learn through the assessment of quality and organization of learning outcomes. *The Curriculum Journal*, 18(2), 195–210. <https://doi.org/10.1080/09585170701446044>
- Csapó, B. (2010). Goals of learning and the organization of knowledge. In E. Klieme, D. Leutner, & M. Kenk (Eds.), *Kompetenzmodellierung. Zwischenbilanz des DFG-Schwerpunktprogramms und Perspektiven des Forschungsansatzes. 56. Beiheft der Zeitschrift für Pädagogik* (pp. 12–27). Weinheim: Beltz.
- Csapó, B., Ainley, J., Bennett, R. E., Latour, T., & Law, N. (2012). Technological issues for computer-based assessment. In P. Griffin, B. McGaw, & E. Care (Eds.), *Assessment and teaching of 21st Century skills* (pp. 143–230). New York, NY: Springer. [https://doi.org/10.1007/978-94-007-2324-5\\_4](https://doi.org/10.1007/978-94-007-2324-5_4)



- Csapó, B., & Csépe, V. (Eds.). (2012). *Framework for diagnostic assessment of reading*. Budapest: Nemzeti Tankönyvkiadó.
- Csapó, B., Lőrincz, A., & Molnár, G. (2012). Innovative assessment technologies in educational games designed for young students. In D. Ifenthaler, D. Eseryel, & X. Ge (Eds.), *Assessment in game-based learning: Foundations, innovations, and perspectives* (pp. 235–254). New York, NY: Springer.
- Csapó, B., Molnár, G., & Nagy, J. (2014). Computer-based assessment of school-readiness and reasoning skills. *Journal of Educational Psychology*, 106(2), 639–650.
- Csapó, B., & Szabó, G. (Eds.). (2012). *Framework for diagnostic assessment of science*. Budapest: Nemzeti Tankönyvkiadó.
- Csapó, B., & Szendrei, M. (Eds.). (2011). *Framework for diagnostic assessment of mathematics*. Budapest: Nemzeti Tankönyvkiadó.
- Csíkós, C., & Csapó, B. (2011). Diagnostic assessment frameworks for mathematics: Theoretical background and practical issues. In B. Csapó & M. Szendrei (Eds.), *Framework for diagnostic assessment of mathematics* (pp. 137–162). Budapest: Nemzeti Tankönyvkiadó.
- Csíkós, C., & Verschaffel, L. (2011). Mathematical literacy and the application of mathematical knowledge. In B. Csapó & M. Szendrei (Eds.), *Framework for diagnostic assessment of mathematics* (pp. 57–93). Budapest: Nemzeti Tankönyvkiadó.
- de Koning, E. (2000). *Inductive reasoning in primary education. Measurement, teaching, transfer*. Zeist: Kerckebosch.
- DeWind, N. K., & Brannon, M. (2016). Significant inter-test reliability across approximate number system assessments. *Frontiers in Psychology*, 7, 310. <https://doi.org/10.3389/fpsyg.2016.00310>
- Dikli, S. (2006). An overview of automated scoring of essays. *The Journal of Technology, Learning, and Assessment*, 5(1). <http://ejournals.bc.edu/ojs/index.php/jtla/article/view/1640/1489>
- Farcot, M., & Latour, T. (2008). An open source and large - scale computer based assessment platform: A real winner. In F. Scheuermann & A. G. Pereira (Eds.), *Towards a research agenda on computer - based assessment: Challenges and needs for European educational measurement* (pp. 64–67). Ispra: European Commission Joint Research Centre.
- Farcot, M., & Latour, T. (2009). Transitioning to Computer-based assessments: A question of costs. In F. Scheuermann & J. Bjornsson (Eds.), *The transition to Computer-based assessment: New approaches to skills assessment and implications for large-scale testing* (pp. 108–116). Brussels: European Communities.
- Frey, A. (2007). Adaptives Testen. In H. Moosbrugger & A. Kelava (Eds.), *Testtheorie und Testkonstruktion* (pp. 261–278). Berlin, Heidelberg: Springer.
- Griffin, P., McGaw, B., & Care, E. (Eds.). (2012). *Assessment and teaching of 21st century skills*. New York: Springer.
- Hu, L., & Bentler, P. M. (1999). Cutoff criteria for fit indexes in covariance structure analysis: Conventional criteria versus new alternatives. *Structural Equation Modeling A Multidisciplinary Journal*, 6(1), 1–55.
- Inhelder, B., & Piaget, J. (1958). *The growth of logical thinking*. London: Routledge and Kegan Paul.
- Jodoin, M., Zenisky, A., & Hambleton, R. K. (2006). Comparison of the psychometric properties of several computer-based test designs for credentialing exams with multiple purposes. *Applied Measurement in Education*, 19(3), 203–220.
- Jurecka, A., & Hartig, J. (2007). Computer- und Netzbasiertes Assessment. In J. Hartig & E. Klieme (Eds.), *Möglichkeiten und Voraussetzungen technologiebasierter Kompetenzdiagnostik* (pp. 37–48). Berlin, Bonn: Bundesministerium für Bildung und Forschung.
- Kettler, R. J. (2011). Computer-based screening for the new modified alternate assessment. *Journal of Psychoeducational Assessment*, 29(1), 3–13.
- Kikis, K. (2010). Reflections on paper-and-pencil tests to eAssessments: Narrow and broadband paths to 21st Century challenges. In F. Scheuermann & J. Bjornsson (Eds.), *The Transition to Computer-based assessment: New approaches to skills assessment and implications for large-scale testing* (pp. 99–103). Brussels: European Communities.
- Klauer, K. J. (1993). *Denktraining für Jugendliche*. Göttingen: Hogrefe.
- Let's Go Learn. Retrieved from <https://frontend.letsgolearn.com/login>

- Martin, R. (2010). Utilising the potential of Computer delivered surveys in assessing scientific Literacy. In F. Scheuermann & J. Bjornsson (Eds.), *The Transition to Computer-based assessment: New approaches to skills assessment and implications for large-scale testing* (pp. 172–177). Brussels: European Communities.
- Martin, R., & Binkley, M. (2009). Gender differences in cognitive tests: A consequence of gender-dependent preferences for specific information presentation formats? In F. Scheuermann & J. Björnsson (Eds.), *The transition to computer-based assessment: New approaches to skills assessment and implications for large-scale testing* (pp. 75–82). Luxembourg: Office for Official Publications of the European Communities.
- Math Garden from the University of Amsterdam. Retrieved from <https://www.mathsgarden.com/more-info/>
- Meijer, R. (2010). Transition to computer-based assessment: Motivations and considerations. In F. Scheuermann & J. Bjornsson (Eds.), *The transition to computer-based assessment: New approaches to skills assessment and implications for large-scale testing* (pp. 104–107). Brussels: European Communities.
- Messick, S. (1995). Standards of validity and the validity of standards in performance assessment. *Educational Measurement: Issues and Practice*, 14(4), 5–8. <https://doi.org/10.1111/j.1745-3992.1995.tb00881.x>
- Mitchell, T., Russel, T., Broomhead, P., & Aldridge, N. (2002). Towards robust computerized marking of free-text responses. In M. Danson (Ed.), *Proceedings of the Sixth International Computer Assisted Assessment Conference*. Loughborough University. Retrieved from [https://dspace.lboro.ac.uk/dspace-jspui/bitstream/2134/1884/1/Mitchell\\_t1.pdf](https://dspace.lboro.ac.uk/dspace-jspui/bitstream/2134/1884/1/Mitchell_t1.pdf)
- Moe, E. (2010). Introducing large-scale computerized assessment – Lessons learned and future challenges. In F. Scheuermann & J. Björnsson (Eds.), *The transition to computer-based assessment: New approaches to skills assessment and implications for large-scale testing* (pp. 51–56). Luxembourg: Office for Official Publications of the European Communities.
- Molnár, G., Greiff, S., Wüstenberg, S., & Fischer, A. (2017). Empirical study of computer based assessment of domain-general dynamic problem solving skills. In B. Csapó, J. Funke, & A. Schleicher (Eds.), *The nature of problem solving* (pp. 123–143). Paris: OECD.
- Molnár, G., & Lőrincz, A. (2012). Innovative assessment technologies: Comparing ‘face-to-face’ and game-based development of thinking skills in classroom settings. In D. Chen (Ed.), *International proceedings of economics development and research. Management and education innovation* (Vol. 37, pp. 150–154). Singapore: IACSIT Press.
- Mullis, I. V. S. (2013). In M. O. Martin (Ed.), *TIMSS 2015 assessment frameworks*. Chestnut Hill, MA: TIMSS & PIRLS International Study Center, Boston College.
- Nguyen, T., Watts, T. W., Duncan, G. J., Clements, D. H., Sarama, J. S., Wolfe, C., & Spitzer, M. E. (2016). Which preschool mathematics competencies are most predictive of fifth grade achievement? *Early Childhood Research Quarterly*, 36, 550–560. <https://doi.org/10.1016/j.ecresq.2016.02.003>
- Nunes, T., & Csapó, B. (2011). Developing and assessing mathematical reasoning. In B. Csapó & M. Szendrei (Eds.), *Framework for diagnostic assessment of mathematics* (pp. 17–56). Budapest: Nemzeti Tankönyvkiadó.
- OECD. (2000). *PISA 2015 assessment and analytical framework: Science, reading, mathematics and financial literacy*. Paris: OECD. <https://doi.org/10.1787/9789264181564-en>
- OECD. (2014). *PISA 2012 results: Creative problem solving: students' skills in tackling real-life problems (volume V)*. Paris: OECD.
- OECD. (2016). *Measuring student knowledge and skills. The PISA 2000 assessment of reading, mathematical and scientific literacy*. Paris: OECD. <https://doi.org/10.1787/9789264255425-en>
- Pachler, N., Daly, C., Mor, Y., & Mellar, H. (2010). Formative e-assessment: Practitioner cases. *Computers and Education*, 54(3), 715–721.
- PAT: Mathematics in New Zealand. Retrieved from <http://www.nzcer.org.nz/tests/pat-mathematics>
- Peak, P. (2005). *Recent trends in comparability studies*. Pearson educational measurement. Retrieved from [http://www.pearsonassessments.com/NR/rdonlyres/5FC04F5A-E79D-45FE-8484-07AACAE2DA75/0/TrendsCompStudies\\_rr0505.pdf](http://www.pearsonassessments.com/NR/rdonlyres/5FC04F5A-E79D-45FE-8484-07AACAE2DA75/0/TrendsCompStudies_rr0505.pdf)

- Pearson (2012). From paper and pencil to computer-based testing (CBT). Retrieved from [http://www.pearsonvue.co.uk/india/Documents/PP\\_to\\_CBT.pdf](http://www.pearsonvue.co.uk/india/Documents/PP_to_CBT.pdf).
- Pearson's MyMathLab. Retrieved from <https://www.pearsonmylabandmastering.com/northamerica/mymathlab/>
- Redecker, C., & Johannessen, Ø. (2013). Changing assessment - towards a new assessment paradigm using ICT. *European Journal of Education, 48*(1), 79–96.
- Ridgway, J., & McCusker, S. (2003). Using computers to assess new educational goals. *Assessment in Education, 10*(3), 309–328.
- Ripley, M. (2010). Transformational Computer-based Testing. In F. Scheuermann & J. Björnsson (Eds.), *The transition to computer-based assessment: New approaches to skills assessment and implications for large-scale testing* (pp. 92–98). Luxembourg: Office for Official Publications of the European Communities.
- SETDA. (2008). National trends report 2008. Enhancing education through technology. Retrieved from <http://www.setda.org/wp-content/uploads/2013/12/National-Trends-Report-2008.pdf>.
- Scheuermann, F., & Björnsson, J. (2009). *The transition to computer-based assessment: New approaches to skills assessment and implications for large-scale testing*. Luxembourg: Office for Official Publications of the European Communities.
- Scheuermann, F., & Pereira, G. A. (Eds.). (2008). *Towards a research Agenda on Computer-based assessment*. Luxembourg: Office for Official Publications of the European Communities.
- Sim, G., & Horton, M. (2005). Performance and attitude of children in Computer based versus paper based testing. In P. Kommers & G. Richards (Eds.), *Proceedings of world conference on educational multimedia, hypermedia and Telecommunications 2005* (pp. 3610–3614). Chesapeake, VA: AACE.
- Strain-Seymour, E., Way, W. D., & Dolan, R. P. (2009). *Strategies and processes for developing innovative items in large-scale assessments*. Iowa City, IA: Pearson Education.
- Szendrei, J., & Szendrei, M. (2011). Scientific and curriculum aspects of teaching and assessing mathematics. In B. Csapó & M. Szendrei (Eds.), *Framework for diagnostic assessment of mathematics* (pp. 95–135). Budapest: Nemzeti Tankönyvkiadó.
- The Diagnostic assessment part of PARCC. Retrieved from [http://futurereadyma.org/uploads/documents/Diagnostics\\_ELA\\_and\\_Math\\_FAQs\\_-\\_December\\_2015.pdf](http://futurereadyma.org/uploads/documents/Diagnostics_ELA_and_Math_FAQs_-_December_2015.pdf)
- Valenti, S., Neri, F., & Cucchiarelli, A. (2003). An overview of current research on automated essay grading. *Journal of Information Technology Education: Research, 2*(1), 319–330.
- Van der Kleij, F. M., Eggen, T. J. H. M., Timmers, C. F., & Veldkamp, B. P. (2012). Effects of feedback in a computer-based assessment for learning. *Computers & Education, 58*(1), 263–272.
- van Lent, G. (2010). Risks and benefits of CBT versus PBT in high-stakes testing. In F. Scheuermann & J. Björnsson (Eds.), *The Transition to Computer-based assessment: New approaches to skills assessment and implications for large-scale testing* (pp. 83–91). Brussels: European Communities.
- Verschaffel, L., De Corte, E., & Lasure, S. (1994). Realistic considerations in mathematical modeling of school arithmetic word problems. *Learning and Instruction, 7*, 339–359. [https://doi.org/10.1016/0959-4752\(94\)90002-7](https://doi.org/10.1016/0959-4752(94)90002-7)
- Watts, T. W., Duncan, G. J., Siegler, R. S., & Davis-Kean, P. E. (2014). What's past is prologue: Relations between early mathematics knowledge and high school achievement. *Educational Researcher, 43*(7), 352–360. <https://doi.org/10.3102/0013189X14553660>