

Identifying national cultures of mathematics education: Analysis of cognitive demands and differential item functioning in TIMSS

Eckhard Klieme

Jürgen Baumert

Max Planck Institute for Human Development, Berlin, Germany

Large-scale assessments of student competencies address rather broad constructs and use parsimonious, unidimensional measurement models. Differential item functioning (DIF) in certain subpopulations usually has been interpreted as error or bias. Recent work in educational measurement, however, assumes that DIF reflects the multidimensionality that is inherent in broad competency constructs and leads to differential achievement profiles. Thus, DIF parameters can be used to identify the relative strengths and weaknesses of certain student subpopulations.

The present paper explores profiles of mathematical competencies in upper secondary students from six countries (Austria, France, Germany, Sweden, Switzerland, the US). DIF analyses are combined with analyses of the cognitive demands of test items based on psychological conceptualisations of mathematical problem solving. Experts judged the cognitive demands of TIMSS test items, and these demand ratings were correlated with DIF parameters. We expected that cultural framings and instructional traditions would lead to specific aspects of mathematical problem solving being fostered in classroom instruction, which should be reflected in differential item functioning in international comparative assessments. Results for the TIMSS mathematics test were in line with expectations about cultural and instructional traditions in mathematics education of the six countries.

Large-scale assessments that serve to monitor educational systems generally follow the logic of domains of knowledge as they are institutionalised in schools. Thus, domains such as reading, writing, mathematics, science, history, or foreign languages are examined. These domains are further structured internally according to curricular or instructional aspects – in terms of arithmetic, algebra and geometry, for example, or reading and listening comprehension, grammar and vocabulary. For reasons of curricular validity, large-scale assessments aim at selecting item samples that will achieve the broadest possible coverage of such subdomains. Therefore, it is now standard practice for large-scale assessments to employ multi-matrix designs. With such designs, students do not work on all items, but each testtaker is presented with only a subset of the item sample. The test's content coverage can thus be substantially expanded, while keeping the testing time constant.

The rationale of large-scale assessments is not to measure theoretically defined psychological constructs, but to evaluate performance in an institutionally defined knowledge domain. Consequently, these assessments capture complex proficiency syndromes which include various interacting psychological abilities and heterogeneous content components. This makes it all the more remarkable that such proficiency syndromes can generally be approximated on a single unidimensional measurement scale. The fit of the unidimensional Rasch model (or any other unidimensional IRT model) is usually almost as good as that of multidimensional models.

The relatively good fit of unidimensional test models is of great benefit for system monitoring purposes and international comparisons because it allows for parsimonious descriptions of results and provides a robust basis for time-series analyses. From the perspective of teaching and learning, however, this parsimony is a drawback, as results provide no information on the specific strengths or weaknesses of different student populations or educational treatments. For teaching and learning to be improved, specific diagnostic information is needed that helps to identify potential points of intervention.

This diagnostic weakness of large-scale assessments forms the starting point for the present article. It is a robust finding that unidimensional IRT models never show a perfect fit in large samples. This misfit is generally regarded as negligible specification error or error variance. The present article investigates this presumed error variance from the perspective of teaching and learning. The underlying assumption is that the misfit can be broken down into a negligible error component and a systematic component reflecting an unmodelled multidimensionality of the test that results from the complexity of individual test items. With this approach, specific strengths and weaknesses of certain student populations, as well as specific effects of different curricular and instructional traditions can be revealed.

Profiles of this kind emerge in so-called differential item functioning (DIF). DIF exists when persons with different group membership but identical overall test scores differ systematically with regard to the probability of solving test items that demand a particular ability or particular prior knowledge. On a technical level, this means that it is possible to identify group-specific item characteristic curves (cf. Camilli & Shepard, 1994; Holland & Wainer, 1993; Scheunemann & Bleistein, 1999).

In the context of diagnosing individuals' achievement, differential item functioning has been interpreted as indicating unwelcome item bias, and as precluding fair comparisons of student outcomes. Items with significant DIF with respect to minority populations are regularly excluded from the Scholastic Aptitude Test (SAT), for example. This type of item bias, which may affect the fairness of intercountry comparisons, also occurs in international comparative large-scale assessments, where it can often be attributed to translation errors (Brislin, 1986; Ercikan, 1998). As large-scale assessments do not aim to evaluate performance of individuals, however, not all cases of DIF necessarily have to be interpreted as item bias that will jeopardise the fairness of the test. Instead, DIF can be viewed as an indicator for differential effects of specific curricular or instructional conditions (Miller & Linn, 1988; Tatsuoka, Linn, Tatsuoka, & Yamamoto, 1988).

Recent publications on assessment share the position that DIF can also provide substantial information that helps to gain a deeper understanding of relative strengths and weaknesses in subpopulations (Calvert, 2001; Keeves & Masters, 1999; Pellegrino, Chudowsky, & Glaser, in press). According to this view, DIF is not necessarily an indicator of bias or methodological flaw, but can "provide information of value in education, because the existence of bias reflects (...) differences in the learning experiences involved for providing a correct response to the item" (Keeves & Masters, 1999, p. 12).

In the present article, we examine DIF in international comparison using data from the advanced mathematics test for the upper secondary sample of the Third International Mathematics and Science Study (TIMSS) (Mullis, Martin, Beaton, Gonzalez, Kelly, & Smith, 1998; Baumert, Bos, & Lehmann, 2000). We work on the underlying assumption that educational systems not only vary in their overall effectiveness, but that they produce different patterns of outcomes. Students in different countries have different learning experiences, which

can lead to country-specific strengths and weaknesses. In using DIF to identify proficiency profiles, we try to explore the implicit multidimensionality of the TIMSS test which is treated as error variance in the unidimensional IRT approach. As van der Linden (1998, p. 574) put it, "in international assessments the achievements are bound to represent multidimensional rather than unidimensional knowledge. Also, national populations can be expected to have different distributions on each of the dimensions; in fact, international assessments are designed just to detect such differences."

Theoretical framework

The issue of country-specific profiles in student outcomes has, in fact, been a concern for international educational research since the very start of comparative student assessment in the 1960s (Husén & Postlethwaite, 1996). However, this research concentrated on curriculum effects on learning outcomes. The straightforward hypothesis was that a country's achievement results will be better in areas that constitute an important part of the country's curriculum and to which a relatively high amount of learning time is devoted. The coverage of topics in the educational system was assessed by calculating an "opportunity to learn" (OTL) index within each school. Mathematics teachers were asked to rate each item in terms of the relative number of students in the school who had been exposed to the mathematical content the item addresses. When aggregating these OTL ratings as well as test scores across countries, it emerged that up to two-thirds of the between-country variance in the IEA's First International Mathematics Study (FIMS) could be explained by differences in OTL (Wolf, 1998, p. 499). These results have now been replicated by a number of other researchers (McDonnell, 1995; Muthén, Huang, Jo, Khoo, Goff, Novak, & Shih, 1995; Westbury, 1993).

In contrast to the OTL approach, our investigation addresses the question of whether profiles in student outcomes can be identified with regard to different cognitive demands of test items. The analyses of the cognitive demands of the test items are carried out within the theoretical framework for solving mathematical word problems developed by Reusser (1996), who elaborated on Kintsch and Greeno's (1985) model of discourse processing and reasoning in working on mathematical word problems. According to Reusser, word problems in mathematics are processed in five consecutive steps. The process starts with the comprehension and, in some cases, a reformulation of the word problem. This is followed by the formulation of an appropriate mathematical model that determines the mathematical operations to be performed. Once these mathematical operations – which are often of a numerical nature – have been applied, the results have to be translated back into the context of the original word problem. Following Freudenthal (1983), Neubrand, Biehler, Blum, Cohors-Fresenborg, Flade, Knoche, Lind, Löding, Möller, and Wynands (2001) describe this sequence as the *complete cycle of mathematical modelling*.

In principle, this model can be generalised to all mathematical tasks. Depending on the type of problem, the different parts of the modelling cycle will be covered to a varying extent and with varying intensity. Because we intend to use a more elaborate form of this model as a heuristic tool for the analysis of the cognitive demands of test items, the individual components of the model will be described in some detail:

- In the first step of the mathematical modelling process, a propositional representation of the problem is constructed on the basis of the information provided. In word problems, this is the text base. Yet, the problem may also be represented visually in the form of a graph or diagram. A propositional understanding of the information contained in such a chart is basically equivalent to the text base.
- In the second step, a more complex situation model is developed on the basis of this propositional representation. Drawing on prior knowledge, experiences, and plausible inferences, the original problem is embedded in a meaningful context. Because

mathematical word problems generally have a narrative structure, Reusser (1996) terms these mental models “episodic problem models”. In our generalised conception, however, this situation model can also involve the systematic embedding of the problem in an argumentative context.

- In a third step, the situation model is transformed into a formal mathematical representation. Reusser describes this step – which primarily involves the reduction and abstraction of information – as mathematisation in the narrower sense. Where demanding mathematics tasks, such as those administered in the advanced TIMSS test, are concerned, however, the relation between steps 2 and 3 will need to be conceptualised in a more elaborate way. We assume that, depending on the difficulty and complexity of the problem, the process of developing a situation model and engaging in mathematisation in the narrower sense may involve several feedback loops. This serves the iterative optimisation of two steps: (a) restructuring the problem situation to ensure the fit with a certain mathematical structure, and (b) formal modelling in the sense of identifying variables, parameters, and structural relations. Lower-level problem solving activities such as planning intermediate steps or working back and forth are of crucial importance here.
- In the fourth step of the problem solving cycle, the relevant operations have to be identified in the formal mathematical model and applied in the correct combination and sequence. This step calls for the application of different kinds and levels of declarative and procedural mathematical knowledge. Following Stein, Grover, and Henningsen (1996), we distinguish between the following:
 - (1) declarative knowledge of facts and procedures,
 - (2) algebraic operations,
 - (3) arithmetical operations, and
 - (4) conceptual understanding.
 Conceptual understanding is defined as understanding the relation between various mathematical concepts, knowing typical examples and counterexamples, being aware of contexts in which the concept may be applied and being able to discriminate the mathematical concept from similar notions that are rooted in everyday experience. The cognitive demands of test items differ depending on which knowledge components are required for the necessary mathematical operations to be performed.
- In the fifth and final step, the results of the mathematisation process and the mathematical operations performed have to be translated back into the context of the original problem and interpreted in a meaningful way. This interpretation must result in a satisfactory answer to the question originally posed in the task.

This extended and generalised model of mathematical problem solving was used as a basis for the development of a system to classify the cognitive demands of mathematical test items. The classification system takes account of both the individual steps of the modelling cycle and the knowledge components implicit in the mathematical operations. Table 1 provides an overview of these process components and cognitive demands. Mathematical problems and test items can be differentiated according to the extent to which and the intensity with which specific demands will become salient.

The basic assumption of our approach is that specific aspects of the generalised model of mathematical problem solving will be accentuated and fostered depending on the cultural context of maths instruction, shared epistemological beliefs about doing and learning mathematics, curricular and instructional traditions, and the rationale underlying the teaching routines that prevail in the classroom. We expect that these instructional patterns are reflected in differential proficiency profiles and that empirical evidence for such profiles will emerge in the form of differential item functioning in international comparative assessments. In other words, strengths and weaknesses of students from different countries should not only result from differences in learning opportunities but also from differences in educational philosophies

and instructional traditions which lead to culture-specific lesson scripts in mathematics classrooms.

Table 1

Process components and cognitive demands in mathematical problem solving

Process component	Cognitive demands
Propositional representation	Text comprehension Processing of visual information
Situational representation mathematisation	Understanding situational contexts Formal modelling Restructuring of problems Problem-solving activities
Inner-mathematical operations (reasoning and calculations)	Declarative knowledge of facts and procedures Algebraic operations Arithmetical operations Conceptual understanding
Interpretation and translation	Interpreting diagrams

Previous research

International comparative research on cultural scripts in math education and their impact on learning outcomes is scarce. The TIMS-Video study (Baumert, Lehmann, Lehrke, Schmitz, Clausen, Hosenfeld, Köller, & Neubrand, 1997; Stigler, Gonzalez, Kawanaka, Knoll, & Serrano, 1996), which investigated instructional practice in Japan, Germany, and the USA, and the study on “characterising pedagogical flow” in mathematics lessons conducted by Schmidt, Jorde, Cogan, Barrier, Gonzalo, Moser, Shimizu, Sawada, Valverde, McKnight, Prawat, Wiley, Raizen, Britton, and Wolfe (1996) in the run-up to TIMSS, are two of the few relevant studies in this field. Characteristic pedagogical flow was defined as “culturally distinct and rationally characteristic patterns in which curriculum and pedagogy intertwined within classrooms, (...) a characteristic interaction between curriculum and pedagogy in lessons. Presumably, this interaction stems from certain national beliefs together with the particular training and experience teachers have had that lead them to share these beliefs” (Cogan & Schmidt, 1999, p. 82). Based on interpretative analyses of classroom observations, the authors drew the following conclusions:

- “French lessons were characterized by formal and complex subject matter that teachers actively organized and presented to students”,
- “the Japanese lessons were characterized as built around a consideration of multiple approaches to carefully chosen practical examples or activities, through which the teacher led students into an understanding of mathematical concepts and relationships”,
- “lessons from Norway were characterized by student activity, both individually and in small groups”,
- while in the US, “both teacher and student activity tended to emphasize the basic definitions and procedures of mathematics” (Cogan & Schmidt, 1999, pp. 79-81).

The observations on instruction in Japan and the USA correspond with the findings by Stigler et al. (1996). Moreover, both Stigler et al. (1996) and Baumert, Lehmann, et al. (1997) describe German mathematics instruction as a sequence of short teacher questions and student answers, ultimately converging in the solution expected by the teacher. This description is consistent with findings on German mathematics instruction (Bauersfeld, 1980; Voigt, 1984; see also Cobb & Bauersfeld, 1995; Seeger, Voigt, & Waschescio, 1998).

None of these studies investigated whether differences in instructional practices are reflected in country-specific achievement profiles, however. Nevertheless, there are a few descriptive results which lend support to the idea that educational cultures not only differ in their overall achievement levels or their achievement profiles in different curricular domains, but also with regard to process-related variables. In comparing the problem-solving behaviour of US and Japanese students, Becker, Sawada, and Shimizu (1999) found that Japanese students not only show a higher level of technical mathematical knowledge, but also exhibit higher levels of sophistication in their problem solving.

Country-specific profiles in the TIMSS lower secondary tests for mathematics and science have been investigated in several studies. Schmidt, Jakwerth, and McKnight (1998) found that country rankings in these tests differed depending on the cognitive demands of items selected for comparison. Ramseier (1999) asked science experts to rate every item from the TIMSS lower secondary science test with regard to cognitive level and knowledge of scientific terminology required. He found that the *relative* item difficulty for Swiss students increased when the reproduction of scientific terms was required, but decreased with cognitive level. For mathematics, Neubrand, Neubrand, and Sibberns (1998) and Blum and Wiegand (1998) found that, compared to the international average, German students often failed to solve items that demand mathematical modelling or the interlinking of basic ideas. Yet, German students showed relative strengths in algorithmic reasoning, single-step operations, and reproduction of factual knowledge.

In our own previous research, we have studied differential item functioning across countries in the TIMSS lower secondary mathematics test (Klieme & Bos, 2000) as well as in the TIMSS upper secondary tests for mathematical and science literacy (Baumert, Bos, & Watermann, 2000a,b; Baumert, Klieme, & Watermann, 1999). Klieme and Bos (2000) explored the link between instructional practice and achievement profiles by comparing Japanese students and German students in grade eight mathematics. Based on analyses of the TIMSS-Video material, they expected that Japanese students would be best prepared to solve high-level, cognitively demanding, inner-mathematical tasks, while German students would be relatively well-prepared to cope with standard tasks embedded in application contexts. An analysis of differential item functioning confirmed this prediction. In a similar vein, Baumert, Klieme, and Watermann (1999) assigned TIMSS upper secondary items addressing mathematics and science literacy to different levels of proficiency. They found that differential item functioning across the seven countries under investigation was highly dependent on the proficiency levels and, hence, on the cognitive demands of the tasks in question. Compared to their peers in Switzerland, France, and Sweden, German students consistently showed weaknesses when mathematical modelling and argumentation were required.

The present study

Based on these findings, we work on the assumption that it will be possible to identify country-specific proficiency profiles for students taking advanced mathematics courses at the upper secondary level. We further assume that these profiles are built up cumulatively across the school career as a result of different philosophies of math education and different instructional scripts. In order to explore these assumptions, the present study compares the results achieved by students from Austria, France, Germany, Sweden, Switzerland, and the United States in the TIMSS advanced mathematics test. The following hypotheses are explored:

- (1) National proficiency profiles can be aptly described with reference to our generalised model of mathematical problem solving.
- (2) The national outcome profiles will correspond with the culture-specific scripts of mathematics instruction identified in previous research.
- (3) The German-speaking countries of Austria, Germany, and Switzerland share a common tradition of mathematics instruction. Although the countries differ

considerably in terms of the overall level of their students' mathematics competence, it will be possible to identify similar patterns of *relative* strengths and weaknesses across these countries.

- (4) For students from German-speaking countries, tasks that entail complex cognitive demands and require problem-solving activities and mathematical argumentation will present a particular challenge.
- (5) French students will show relative strengths in test tasks requiring complex subject matter knowledge.
- (6) Swedish students will be particularly well-prepared to tackle open problems requiring a situational understanding, restructuring and formal modelling.
- (7) US students will be relatively good at procedural tasks.

Method

Participants

The TIMSS advanced mathematics test was administered in 18 countries to representative samples of students attending a pre-university mathematics course in the final year of upper secondary education in the 1995/96 academic year. Six of these countries were selected for the present study. In *Austria*, a sample of $N=599$ students was selected. The sample is representative for the target population of all students attending a mathematics course in the final year of general upper secondary school or high-level vocational school. In *France*, the sample comprised $N=796$ students in the 12th grade of the *Lycée d'Enseignement Générale Scientifique*. The target population consisted of all students attending the 12th grade in the natural science track. In *Germany*, a sample of $N=2,189$ students was drawn from the target population of *Gymnasium* track students attending either a basic or an advanced course in mathematics in the final grade. In *Sweden*, $N=749$ students were sampled from the target population of 12th graders who had opted for the natural science or technology track in upper secondary school. In *Switzerland*, the sample consisted of $N=1,072$ students from the final year of all types of *Gymnasium*. The sample in the *United States*, finally, comprised $N=2,349$ students drawn from the target population of all 12th graders attending a calculus course, a pre-calculus course, or an advanced placement course in calculus.

Measures

The TIMSS advanced mathematics test consisted of 66 items, administered in a multi-matrix sampling design. According to the authors of the test, "the advanced mathematics test reflected current thinking and priorities in the field of mathematics" (Mullis et al., 1998, p. B-7), that is, the items cover the content and cognitive demands of pre-university curricula. Mathematics experts classified 40 percent of the advanced mathematics tasks as application or problem-solving items, while 60 percent were judged to require the knowledge or use of more or less complex procedures. Most of the items from the mathematics test have since been released to the public on the Internet site of Boston College (see also Baumert, Bos, Klieme, Lehmann, Lehrke, Hosenfeld, Neubrand, & Watermann, 1999). The test was subjected to uni-dimensional scaling on the basis of the multinomial Rasch model (Adams & Wilson, 1996).

Procedure

The analytical procedure implemented in the present study comprised the following three steps:

- Step I:* Analysis of the cognitive demands of the TIMSS advanced mathematics items on the basis of expert ratings.
- Step II:* Estimation of parameters for differential item functioning in the five selected countries.
- Step III:* Computation of correlations of demand ratings and DIF-parameters to identify relative strengths and weaknesses with regard to specific demands.

For Step I, a system for the classification of cognitive demands of mathematics items was developed, based on the generalised model of mathematical problem solving described above. The first version of the rating system was tested on all TIMSS items by six experts in the field of mathematics education. Based on feedback from the experts, interrater reliability results, and correlations between the ratings and the Rasch threshold parameters, the rating system was revised. The final version was then applied independently by ten new mathematics experts. All experts were university staff with extensive teaching experience at the secondary level as well as research experience in the field of mathematics education. The experts received verbal descriptions of the components of the classification system, but there was no interactive training session.

The experts were asked to evaluate each of the 66 mathematics items on the basis of the classification system presented in Table 1. They rated the importance of each cognitive demand for solving each of the test items on a four-point Likert scale ranging from 0 (= the demand is of no importance at all for the solution of the item) to 3 (= the demand is crucial for success in solving the item). In the final version of the rating system, a further rating dimension – termed *curricular level* – was added to the original set of cognitive demands. Here, experts were asked to judge on a three-point scale whether the mathematical knowledge needed to solve the item was elementary knowledge as gained from everyday experience or primary education (Level 0), knowledge typically acquired in lower secondary school (Level 1), or knowledge specifically from the upper secondary advanced mathematics curriculum (Level 2). Generalisability coefficients (Shavelson & Webb, 1991) were computed in order to test the reliability of the ratings. These coefficients, which are presented in Table 2, range from $\rho = .16$ (formal modelling) to $\rho = .62$ (interpreting diagrams).

Table 2

Cognitive demands of TIMSS advanced mathematics items as rated by German experts

Demand	Generalisability (10 experts)	Number of items with mean ratings > 1.5	Correlation with item difficulty
Text comprehension	.40	6	.08
Processing of visual information	.49	14	.09
Understanding situational contexts	.49	2	-.02
Formal modelling	.16	12	.35**
Restructuring of problems	.26	2	.13
Problem-solving activities	.25	7	.34**
Curricular level	.61	32	.44***
Declarative knowledge of facts and procedures	.37	56	.53***
Algebraic operations	.49	26	.51***
Arithmetical operations	.47	30	.00
Conceptual understanding	.27	24	.54***
Interpreting diagrams	.62	9	-.09

Note. ** $p < .01$; *** $p < .001$.

Judged by the standard criteria of generalisability theory (Shavelson & Webb, 1991), these indices are low but sufficient because ratings for single items will not be interpreted. In fact, we are only interested in correlations between cognitive demands and DIF across items. The process component that was least reliably judged by the experts is the mathematisation component. This finding may reflect the fact that the mathematisation process is not yet clearly enough defined in theories of mathematical thinking and learning.

For Step II, the advanced mathematics test from TIMSS was scaled using the multinomial Rasch model developed by Adams and Wilson (1996) and implemented in their ConQuest software (Wu, Adams, & Wilson, 1998). This IRT (Item Response Theory) model basically assumes that the item characteristic curves (ICCs) for all items within a test have a similar shape which can be described by a logistic function as depicted in Figure 1. Each item is uniquely identified by its difficulty parameter, indicating the point on the ability/difficulty scale where the probability of solving the item correctly is .50. This one-parameter IRT model (Rasch model) is used routinely in international comparative assessments and has successfully been applied to all items in the TIMS study. Item-fit indices showed acceptable fit across countries. Furthermore, a reanalysis of the data by the German TIMSS research group (Klieme, 2000) revealed that, in principle, the tests for advanced mathematics can be appropriately described as unidimensional. The model fit is not perfect, however, meaning that it is worth testing for unmodelled multidimensionality. When subdimensions of the mathematics test are defined in terms of content domains or cognitive demands, moreover, the intercorrelations – after correcting for attenuation – range between $r=.77$ and $.87$, suggesting that the relationships among items entailing different demands are not perfect.

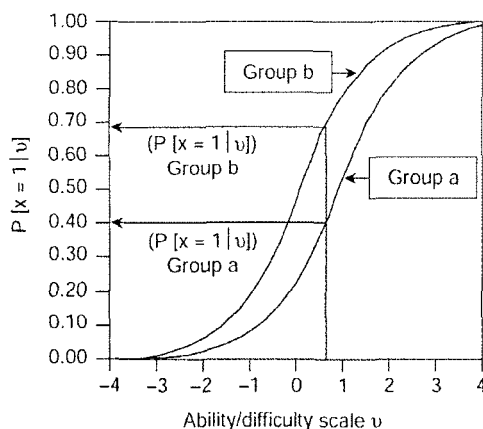



Figure 1. Differential item functioning in two groups (a and b)

Thus, it seems reasonable to expect differential item functioning (DIF) between countries. The basic idea behind the DIF-approach is illustrated in Figure 1. If there is differential item functioning between two groups a and b, the item characteristic curves (ICCs) will differ when calculated separately for each group. One group is called the reference group, the other the focus group. In our analyses, the German student sample is always used as the reference group, while the focus group consists either of the student sample from a specific comparison country (Austria, Switzerland, France, Sweden, or the US) or the combination of all five comparison samples.

A separate Rasch analysis was performed for each focus sample. The item-difficulty parameters estimated within each model were decomposed into four components (see Figure 2): the general mean (which is an arbitrary value obtained by averaging across items and countries),

the effect for the focus country (which indicates the overall performance level of that country's students in relation to German students), the effect for the item (which describes the overall difficulty of that item compared to other questions from the TIMSS test), and the so-called DIF-parameter (i.e., the component describing the strength of the item-by-country interaction). The DIF-parameter was calculated such that positive values reflect relative strengths of students from the focus country, while negative values indicate that German students perform relatively well on that particular item. The average of the DIF-parameters across the items of a test is 0.

$$\text{Item difficulty} = \text{General mean (reference country and focus country)} + \text{Effect for focus country} + \text{Effect for item} + \text{Effect for item-by-country interaction}$$



DIF-parameter

Figure 2. Decomposition of the IRT item difficulty parameter in DIF analysis

In Step III, the DIF-parameters were correlated with the aggregated expert ratings of cognitive demands. A positive correlation indicates that the more important the respective cognitive demand is for solving an item, the more pronounced is the relative strength of students from the focus country compared to German students.

Results

Structure of the TIMSS advanced mathematics tests in terms of cognitive demands

In addition to the generalisability coefficients for the cognitive demand ratings, Table 2 provides basic information on two aspects of the structure of the TIMSS mathematics test. The second column shows for how many items experts judged each demand dimension to be important in achieving a correct answer. A mean rating of at least 1.5 was set as a critical threshold, indicating that the demand was judged to be of intermediate importance. The third column shows the correlation between the mean cognitive demand ratings and the item-difficulty indices of the Rasch model.

The expert ratings indicate that the TIMSS advanced mathematics test has two main foci. A substantial number of the items entail cognitive demands that are of particular relevance in the inner-mathematical phase of solving the problem. These items demand declarative knowledge of facts and procedures, conceptual understanding, and algebraic operations. At the same time, items involving such cognitive demands proved to be particularly difficult on the international level; the correlations between these demands and item difficulty are substantial ($r > .50$). As expected, tasks involving the typical curricular level of pre-university mathematics courses proved to be more difficult than tasks referring to material typically covered already in lower secondary school ($r = .44$). A second main focus of the test was identified in the form of items involving the transformation of situation models into formalised mathematical expressions. This transformation occurs in Step 3 of our theoretical model. The relevant items involve particular demands with respect to formal modelling and require problem-solving activities. These items, too, tended to be more difficult on the international level ($r = .35/.34$). Table 2 also shows that the TIMSS test contains a small, but noteworthy number of items that make particular demands with respect to propositional and situational understanding. These items are distributed relatively evenly across the difficulty scale ($r = .08/.09$ and $-.02$).

In order to provide a more detailed insight into the demand configurations on the item level, a cluster analysis was carried out to identify homogeneous groups of mathematic items on the basis of their mean demand ratings. This cluster analysis was performed using the Ward algorithm and Euclidean distances. In the context of our generalised model of mathematical problem solving, a five-cluster solution proved to be the most appropriate. Table 3 describes the centers of these five clusters in terms of average ratings on the cognitive demand scales.

Table 3

Final cluster centers identified on the basis of averaged expert ratings

Cognitive demands	High level mathematisation and problem solving (<i>N</i> =12)	Understanding and interpreting diagrams (<i>N</i> =11)	Solving word problems (<i>N</i> =13)	Inner- mathematical reasoning and calculations (<i>N</i> =27)	Propositional representation and reasoning (<i>N</i> =3)
Text comprehension	.86	.68	1.28	.30	1.93
Processing of visual information	1.82	1.04	.41	.50	1.17
Understanding situational contexts	.17	.31	1.03	.00	.70
Formal modelling	1.03	1.06	1.26	.96	.37
Restructuring of problems	1.20	.61	.69	.48	1.07
Problem-solving activities	1.26	.54	.88	.63	1.13
Curricular level	1.31	1.38	1.50	1.59	.47
Declarative knowledge of facts and procedures	2.03	1.78	2.08	2.07	.33
Algebraic operations	.86	.18	.91	1.70	.13
Arithmetical operations	.86	.36	1.47	1.71	.40
Conceptual understanding	1.45	1.27	1.73	1.01	.17
Interpreting diagrams	.19	2.00	.16	.18	.00
Mean item difficulty ¹	633	585	613	562	473

Note. ¹Standardised TIMSS-Scale: $\alpha=500$; $sd=100$.

With 27 of the 66 tasks, cluster 4 is the largest of the five clusters. It contains tasks that primarily demand inner-mathematical reasoning and calculation. Cluster 5, which includes only three tasks and thus presents the smallest of the clusters, forms a complement to this: The tasks in this cluster require predominantly the construction of a propositional text base and a few steps in the mathematisation process. Inner-mathematical operations are of no importance here, as these tasks primarily call for logical thinking and text comprehension.

In contrast to clusters 4 and 5, clusters 1 to 3 are mixed types, that is, they contain tasks that involve demands associated with all components of the theoretical problem solving model, albeit with varying levels of intensity. The three clusters are medium in size (12, 11, and 13 items respectively) and can be distinguished as follows: Cluster 3 mainly contains classical "word problems". In comparison to the other clusters, understanding of situational contexts plays a more central role, as do demands related to transforming a situation model into a formal expression and inner-mathematical operations. Cluster 2 is the only cluster in which the components of interpreting and translating the solution back to a verbal context come to the fore. These tasks do not entail any algebraic demands, but call for the interpretation of diagrams. Tasks in Cluster 1, finally, demand mathematisation and problem solving on a higher curricular and inner-mathematical level. The majority of these tasks involve geometry.

The five-cluster solution is highly compatible with our theoretical model. The centroids of the clusters reflect the degree to which individual components are salient in the solution of the different test tasks. Furthermore, an analysis of variance revealed that the items in the individual clusters differ markedly in terms of their difficulty levels. The most difficult tasks

are those in Cluster 1, which demand mathematisation in the narrower sense and, at the same time, inner-mathematical knowledge and understanding. The next most difficult items are those in Cluster 3, which require the complete modelling process to be applied. In comparison, tasks requiring inner-mathematical operations only (Cluster 4) are much easier.

The TIMSS advanced mathematics test is traditional in the sense that the majority of the tasks stress inner-mathematical operations. As such, the test corresponds with widespread habits in mathematics instruction. However, the TIMSS test also contains a series of tasks that make particularly high demands with respect to the formulation of a mathematical model or that require students to carry out almost the entire modelling cycle. It is precisely these tasks that make the TIMSS test particularly suitable for an attempt to identify differential proficiency profiles in international comparison.

Country-specific strengths and weaknesses

In line with our expectations, comparisons of the five selected countries show that students from these countries differ with regard to the probability of solving certain items, even when one controls for differences in overall performance. Particularly when Germany is compared with Sweden and France, the DIF-parameters reach substantial levels ($DIF > 1.0$). With values on the Logit-Scale of 10 and higher, this means that the difference in the mean performance of German and Swedish students – which is already considerable – would increase by more than one standard deviation if the TIMSS advanced mathematics test consisted only of items with DIF-scores of this magnitude. For the interpretation of the DIF-parameters, however, it is essential to know whether these parameters can be systematically linked to curricular characteristics or – as assumed in the present paper – to the cognitive demands of the items. Only under this condition can the DIF-parameters be interpreted as indicators for unmodelled multidimensionality. The present study tests for such systematic relations by examining the correlations between cognitive demands and DIF-parameters. The results are presented in Table 4.

Table 4

*Relative national strengths and weaknesses in advanced mathematics, compared to Germany**

Demand	Correlation between cognitive demand and DIF in favour of					
	Austria	Switzerland	France	Sweden	USA	All
Text comprehension			.35		.38	.29
Processing of visual information			-.45	-.25	-.46	-.32
Understanding situational contexts				.32		
Formal modelling				.22		.21
Restructuring of problems			-.40	.24		
Problem-solving activities			-.22	.22		.20
Curricular level			.35	n.a.	.38	.29
Declarative knowledge of facts and procedures	.22	.31	.28		.32	.38
Algebraic operations			.29	.36	.34	.28
Arithmetical operations				.34	.28	.33
Conceptual understanding	.35			.25	.25	.37
Interpreting diagrams						-.22

Note. * Correlations presented in the table are statistically significant; $p < .05$.

The pattern of results allows for a straightforward interpretation of relative strengths and weaknesses: As hypothesised by Schmidt et al. (1996), the French culture of mathematics education seems to be unique in placing a strong emphasis on pure mathematical reasoning

and high-level knowledge, while neglecting, to some extent, the aspect of constructing mathematical models for extra-mathematical situations. This confirms the expectation that the tradition of the French *Lycée d'Enseignement Générale Scientifique* has a strong focus on teaching mathematics from a systematically developed inner-mathematical perspective.

Judged by international comparative standards (Mullis et al., 1998), the overall performance level of Swedish students is as high as that of French students. Yet, the Swedish culture of mathematics education represents the Scandinavian tradition of application-oriented instruction, which, in a way, presents a complement to the French pattern. The strengths of the Swedish students clearly lie in the building of mental models and mathematising processes. However, compared to German students, the Swedish TIMSS participants are also quite strong in algebraic and arithmetical operations.

Students from the US – relative to their overall performance level, which is much lower than that of the French and Swedish students – show strengths in cognitive demands that relate to mathematical reasoning processes. This seems to be a result of an approach to instruction which focuses on declarative and procedural knowledge and ease of mathematical operations, as opposed to in-depth understanding, application, and applied problem solving. This instructional pattern identified by Stigler et al. (1996) in the TIMS-Video Study involving students from the lower secondary level seems to persist even in the upper secondary level.

The three German-speaking countries (Austria, Switzerland, and Germany) seem to have quite similar profiles, that is, there are almost no significant correlations between cognitive demands and DIF-parameter estimates for Austria and Switzerland. Although students from Switzerland in particular show a much higher overall performance than students from Germany, the shared pedagogical traditions of the three countries seem to result in similar profiles. The only areas in which German students seem to show relative strengths compared to students from other countries are the processing of visual information and the interpretation of diagrams. In other words, the German culture of mathematics education seems to foster the use of visual representations, but it is relatively weak when it comes to promoting the core components of mathematical reasoning.

Summary and discussion

National profiles in learning outcomes which may, in turn, be interpreted as differential effects of cultural backgrounds and educational traditions have played an important role in large-scale international surveys. However, many of the attempts to examine such profiles involved two major restrictions: Methodologically, they were most often based on calculations of percent-correct indices for single items – a method that is inappropriate when judged from the perspective of recent psychometric insights (van der Linden, 1998). From a substantive point of view, moreover, most studies have restricted their investigations to exploring how differences between national curricula with regard to content coverage result in differential profiles.

The approach taken in the present study aims at identifying *cognitive* profiles of students from different countries. Based on a generalised model of mathematical problem solving, characteristic cognitive demands were defined for each step involved in the process of solving mathematics items. More specifically, the model formed the basis for the development of a classification system that mathematics experts used to evaluate the items from the TIMSS advanced mathematics test. The expert ratings were subsequently correlated with IRT parameters describing the differential item functioning across countries (DIF-parameters). Thus, profiles of relative strengths and weaknesses could be described for each country. The starting point for this approach is the robust finding that the Rasch scaling model holds only approximately across subpopulations of large-scale studies. Observed deviations from the Rasch model, which have mostly been interpreted as error or bias in previous research, can – at least partially – be interpreted substantively when it is combined with in-depth task analyses.

In applying this methodology to international comparative data from the TIMS study on advanced mathematics, we were able to demonstrate its feasibility. The results of our analyses were in line with expectations derived from curriculum analyses, case studies, video studies, and expert knowledge in the field of mathematics education. More specifically, our findings support the hypothesis that upper secondary pre-college mathematics education

- in the US focuses on declarative and procedural knowledge,
- in France emphasises knowledge of advanced mathematical concepts while neglecting, to some extent, the development of problem-solving strategies,
- in Sweden is oriented towards application and problem-solving strategies,
- in Germany is weak on advanced knowledge and understanding, with strengths only in the use of visual or graphical representations, and
- in the three German-speaking countries shows a similar profile, although Swiss students perform at a much higher level overall than Austrian and German students do.

In the long run, progress in the area of proficiency profile analysis, particularly in international comparative studies, will be dependent on the improvement of multidimensional Rasch models which permit the inherent multidimensionality of test items to be explicitly modelled. Multidimensional Rasch scaling that takes the factorial complexity of items into account would increase the practical relevance of large-scale assessments substantially. The first steps in this direction have been taken by Adams, Wilson, and Wang (1997) as well as Rost and Carstensen (2000). Hopefully, these steps will be perfected in the near future such that analyses of within-item multidimensionality will become standard procedure in large-scale assessments.

References

- Adams, R.J., & Wilson, M.R. (1996). A random coefficients multinomial logit: A generalized approach to fitting Rasch models. In G. Engelhard & M. Wilson (Eds.), *Objective measurement III: Theory into practice* (pp. 143-166). Norwood, NJ: Ablex.
- Adams, R., Wilson, M., & Wang, W.C. (1997). The multidimensional random coefficients multinomial logit model. *Applied Psychological Measurement*, 21, 1-23.
- Bauersfeld, H. (1980). Hidden dimensions in the so-called reality of a mathematics classroom. *Educational Studies in Mathematics*, 11, 23-29.
- Baumert, J., Bos, W., Klieme, E., Lehmann, R.H., Lehrke, M., Hosenfeld, I., Neubrand, J., & Watermann, R. (Eds.). (1999). *Testaufgaben zu TIMSS/III. Mathematisch-naturwissenschaftliche Grundbildung und voruniversitäre Mathematik und Physik der Abschlussklassen der Sekundarstufe II (Population 3)*. Berlin: Max-Planck-Institut für Bildungsforschung (Materialien aus der Bildungsforschung, 62).
- Baumert, J., Bos, W., & Lehmann, R. (Eds.). (2000). *TIMSS/III. Dritte Internationale Mathematik- und Naturwissenschaftsstudie – Mathematische und naturwissenschaftliche Bildung am Ende der Schullaufbahn: Vol. 2. Mathematische und physikalische Kompetenzen am Ende der gymnasialen Oberstufe*. Opladen: Leske + Budrich.
- Baumert, J., Lehmann, R., Lehrke, M., Schmitz, B., Clausen M., Hosenfeld, I., Köller, O., & Neubrand, J. (1997). *TIMSS – Mathematisch-naturwissenschaftlicher Unterricht im internationalen Vergleich. Deskriptive Befunde*. Opladen: Leske + Budrich.
- Baumert, J., Bos, W., & Watermann, R. (2000a). Mathematische und naturwissenschaftliche Grundbildung im internationalen Vergleich. In J. Baumert, W. Bos, & R. Lehmann (Eds.), *TIMSS/III. Dritte Internationale Mathematik- und Naturwissenschaftsstudie – Mathematische und naturwissenschaftliche Bildung am Ende der Schullaufbahn: Vol. 1. Mathematische und naturwissenschaftliche Grundbildung am Ende der Pflichtschulzeit* (pp. 135-198). Opladen: Leske + Budrich.
- Baumert, J., Bos, W., & Watermann, R. (2000b). Fachleistungen im voruniversitären Mathematik- und Physikunterricht im internationalen Vergleich. In J. Baumert, W. Bos, & R. Lehmann (Eds.), *TIMSS/III. Dritte Internationale*

Mathematik- und Naturwissenschaftsstudie – Mathematische und naturwissenschaftliche Bildung am Ende der Schullaufbahn: Vol. 2. Mathematische und physikalische Kompetenzen am Ende der gymnasialen Oberstufe (pp. 129-180). Opladen: Leske + Budrich.

- Baumert, J., Klieme, E., & Watermann, R. (1999). Jenseits von Gesamttest- und Untertestwerten: Analyse differentieller Itemfunktionen am Beispiel des mathematischen Grundbildungstests der Dritten Internationalen Mathematik- und Naturwissenschaftsstudie der IEA (TIMSS). In H.-J. Herber & F. Hofmann (Eds.), *Schulpädagogik und Lehrerbildung. Festschrift zum 60. Geburtstag von Josef Thonhauser* (pp. 301-324). Innsbruck: Studien Verlag.
- Becker, J.P., Sawada, T., & Shimizu, Y. (1999). Some findings of the US-Japan cross-cultural research on students' problem-solving behaviours. In G. Kaiser, E. Luna, & I. Huntley. (Eds.), *International comparison in mathematics education* (p. 121-139). London: Falmer Press.
- Blum, W., & Wiegand, B. (1998). Wie kommen die deutschen TIMSS-Ergebnisse zustande? In W. Blum & M. Neubrand (Eds.), *TIMSS und der Mathematikunterricht* (pp. 28-34). Hannover: Schroedel.
- Brislin, R.W. (1985). The wording and translation of research instruments. In W. J. Lonner & J. W. Berry (Eds.), *Field methods in cross-cultural research. Cross-cultural research and methodology series* (vol. 8, pp. 137-164). Beverly Hills, CA: Sage.
- Calvert, T. (2001). *Exploring differential item functioning (DIF) with the Rasch model: A cross-country comparison of gender differences on eighth grade science items*. Seattle, WA: AERA.
- Camilli, G., & Shepard, L.A. (1994). *Methods for identifying biased test items* (vol. 4). Thousand Oaks: Sage.
- Cobb, P., & Bauersfeld, H. (1995). *The emergence of mathematical meaning: Interaction in classroom cultures*. Hillsdale, NJ: Lawrence Erlbaum.
- Cogan, L.S., & Schmidt, W.H. (1999). An examination of instructional practices in six countries. In G. Kaiser, E. Luna, & I. Huntley (Eds.), *International comparison in mathematics education* (pp. 68- 85). London: Falmer Press.
- Ercikan, K. (1998). Translation effects in international assessments. *International Journal of Educational Research*, 29, 543-553.
- Freudenthal, H. (1983). *Didactical phenomenology of mathematical structures*. Dordrecht: Riedel.
- Holland, P.W., & Wainer, H. (Eds.). (1993). *Differential item functioning*. Hillsdale, NJ: Erlbaum.
- Husén, T., & Postlethwaite, T.N. (1996). A brief history of the International Association for the Evaluation of Educational Achievement (IEA). *Assessment in Education*, 3, 129-141.
- Keeves, J.P., & Masters, G.N. (1999). Introduction. In G.N. Masters & J.P. Keeves (Eds.), *Advances in measurement in educational research and assessment* (pp. 1-19). Oxford: Pergamon.
- Kintsch, W., & Greeno, J.G. (1985). Understanding and solving word arithmetic problems. *Psychological Review*, 92, 109-129.
- Klieme, E. (2000). Fachleistungen im voruniversitären Mathematik- und Physikunterricht: Theoretische Grundlagen, Kompetenzstufen und Unterrichtsschwerpunkte. In J. Baumert, W. Bos, & R. Lehmann (Eds.), *TIMSS/III. Dritte Internationale Mathematik- und Naturwissenschaftsstudie – Mathematische und naturwissenschaftliche Bildung am Ende der Schullaufbahn: Vol. 2. Mathematische und physikalische Kompetenzen am Ende der gymnasialen Oberstufe* (pp. 57-128). Opladen: Leske + Budrich.
- Klieme, E., & Bos, W. (2000). Mathematikleistung und mathematischer Unterricht in Deutschland und Japan: Triangulation qualitativer und quantitativer Analysen am Beispiel der TIMS-Studie. *Zeitschrift für Erziehungswissenschaft*, 3.
- McDonnell, L.M. (1995). Opportunity to learn as a research concept as a policy instrument. *Educational Evaluation and Policy Analyses*, 17, 305-322.
- Miller, M.D., & Linn, R.L. (1988). Invariance of item characteristic functions with variations in instructional coverage. *Journal of Educational Measurement*, 25, 205-220.
- Mullis, I.V.S., Martin, M.O., Beaton, A.E., Gonzalez, E.J., Kelly, D.L., & Smith, T.A. (1998). *Mathematics and science achievement in the final year of secondary school. IAE's Third International Mathematics and Science Study*. Chestnut Hill, MA: Boston College.
- Muthen, B., Huang, L., Jo, B., Khoo, S., Goff, G., Novak, J., & Shih, J. (1995). Opportunity-to-learn effects on achievement: Analytical aspects. *Educational Evaluation and Policy Analysis*, 17, 371-403.

- Neubrand, J., Neubrand, M., & Sibbers, H. (1998). Die TIMSS-Aufgaben aus mathematik-didaktischer Sicht: Stärken und Defizite deutscher Schülerinnen und Schüler. In W. Blum & M. Neubrand (Eds.), *TIMSS und der Mathematikunterricht* (pp. 17-24). Hannover: Schroedel.
- Neubrand, M., Biehler, R., Blum, W., Cohors-Fresenborg, E., Flade, L., Knoche, N., Lind, D., Löding, W., Möller, G., & Wynands, A. (2001). Grundlagen der Ergänzung des internationalen PISA-Mathematik-Tests in der deutschen Zusatzhebung. *Zentralblatt für Didaktik der Mathematik*, 33, 1-15.
- Pellegrino, J., Chudowsky, N., & Glaser, R. (Eds.). (in press). *Knowing what students know. The science and design of educational assessment*. Washington, DC: National Academy Press.
- Ramseier, E. (1999). Task difficulty and curricular priorities in science: Analysis of typical features of the Swiss performance in TIMSS. *Educational Research and Evaluation*, 5, 105-126.
- Reusser, K. (1996). From cognitive modeling to the design of pedagogical tools. In S. Vosniadou, E. De Corte, R. Glaser, & H. Mandl (Eds.), *International perspectives on the design of technology-supported learning environments* (pp. 81-103). Mahwah, NJ: Lawrence Erlbaum Ass. Publishers.
- Rost, J., & Carstensen, C. H. (2000). Multidimensional Rasch measurement via item component models and faceted designs. Accepted to *Applied psychological Measurement*.
- Scheuneman, J.D., & Bleistein, C.A. (1999). Item bias. In G.N. Masters & J.P. Keeves (Eds.), *Advances in measurement in educational research and assessment* (pp. 220-234). Oxford: Pergamon.
- Schmidt, W.H., Jakwerth, P.M., & McKnight, C. (1998). Curriculum sensitive assessment: Content does make difference. *International Journal of Educational Research*, 29, 503-527.
- Schmidt, W.H., Jorde, D., Cogan, L.S., Barrier, E., Gonzalo, I., Moser, U., Shimizu, K., Sawada, T., Valverde, G.A., McKnight, C., Prawat, R.S., Wiley, D.E., Raizen, S.A., Britton, E.D., & Wolfe, R.G. (1996). *Characterizing pedagogical flow. An investigation of mathematics and science teaching in six countries*. Dordrecht: Kluwer.
- Seeger, F., Voigt, J., & Waschescio, U. (Eds.). (1998). *The culture of the mathematics classroom*. Cambridge, UK: University Press.
- Shavelson, R.J., & Webb, N.M. (1991). *Generalizability theory: A primer*. Newbury Park, CA: Sage.
- Stein, M.K., Grover, B.W., & Henningsen, M. (1996). Building student capacity for mathematical thinking and reasoning: An analysis of mathematical tasks used in reform classrooms. *American Educational Research Journal*, 32, 455-488.
- Stigler, J.W., Gonzalez, P., Kawanaka, T., Knoll, S., & Serrano, A. (1996). *The TIMSS videotape classroom study: Methods and preliminary findings*. Prepared for the National Center for Education Statistics, U.S. Department of Education, Los Angeles, CA.
- Tatsuoka, K.K., Linn, R.L., Tatsuoka, M.M., & Yamamoto, K. (1988). Differential item functioning resulting from use of different solution strategies. *Journal of Educational Measurement*, 25, 301-319.
- Van der Linden, W.J. (1998). A discussion of some methodological issues in international assessments. *International Journal of Educational Research*, 29, 569-577.
- Voigt, J. (1984). *Interaktionsmuster und Routinen im Mathematikunterricht* [Patterns of interaction and routines in mathematics classrooms]. Weinheim: Beltz Verlag.
- Westbury, I. (1993). American and Japanese achievement... again. *Educational Researcher*, 22, 21-25.
- Wolf, R.M. (1998). Validity issues in international assessments. *International Journal of Educational Research*, 29, 491-501.
- Wu, M.L., Adams, R.J., & Wilson, M.R. (1998). *ACER Conquest. Generalised Item Response Modelling Software*. Unpublished manual, Camberwell, Melbourne, Victoria: Australian Council for Educational Research.

Les évaluations à large échelle concernant les compétences d'étudiants traitent des dimensions assez globales et utilisent des modèles de mesure restreints et unidimensionnels.

Le "differential item functioning", utilisé pour certaines sous-populations, a été interprété comme erreur ou biais. De travaux récents dans le domaine de l'évaluation en éducation laissent supposer cependant que le DIF reflète la multidimensionalité inhérente aux dimensions de compétence et ce qui nous amène à des profils de compétence différentiels. En conséquence, les paramètres des analyses DIF sont aptes à identifier les forces et les faiblesses relatives de certaines sous-populations étudiantes.

Cet article examine les profils de compétences mathématiques chez des étudiants du deuxième cycle de six pays différents (Autriche, France, Allemagne, Suède, Suisse et Etats-Unis). Les analyses DIF ont été combinées avec l'analyse des exigences cognitives des items, basée sur des concepts psychologiques de la résolution de problèmes mathématiques. Des experts ont jugé les exigences cognitives des items TIMSS, ensuite ces jugements ont été mis en rapport avec les paramètres DIF.

Selon notre hypothèse que les différents cadres culturels et traditions d'enseignement devraient se traduire dans une différence priorisée à différents aspects de la résolution de problèmes en classe, phénomène qui devrait se retrouver, en utilisant des analyses DIF, dans les estimations comparatives internationales. Les résultats du test mathématique de TIMSS étaient en accord avec les attentes liées aux traditions culturelles et d'enseignement dans l'enseignement des mathématiques dans les six pays examinés.

Key words: Comparative research, Educational assessment, Mathematical problem solving.

Received: June 2001

Eckhard Klieme. Max Planck Institute for Human Development, Lentzeallee 94, D-14195 Berlin, Germany, klieme@mpib-berlin.mpg.de

Current theme of research:

School effectiveness. Instructional quality. Problem solving in mathematics and science. Large-scale assessment.

Most relevant publications in the field of Psychology of Education:

Klieme, E. (1989). *Mathematisches Problemlösen als Testleistung*. Frankfurt a.M.: Lang.

Klieme, E. (2000). Fachleistungen im voruniversitären Mathematik- und Physikunterricht: Theoretische Grundlagen, Kompetenzstufen und Unterrichtsschwerpunkte. In J. Baumert, W. Bos, & R. Lehmann (Eds.), *TIMSS/III. Dritte Internationale Mathematik- und Naturwissenschaftsstudie – Mathematische und naturwissenschaftliche Bildung am Ende der Schullaufbahn: Vol. 2. Mathematische und physikalische Kompetenzen am Ende der gymnasialen Oberstufe* (pp. 57-72). Opladen: Leske + Budrich.

Klieme, E., & Bos, W. (2000). Mathematikleistungen und mathematischer Unterricht in Deutschland und Japan. Triangulation qualitativer und quantitativer Forschungsansätze im Rahmen der TIMS-Studie. *Zeitschrift für Erziehungswissenschaft*, 3, 359-379.

Klieme, E., Funke, J., Leutner, D., Reimann, P., & Wirth, J. (2001). Problemlösen als fächerübergreifende Kompetenz. Konzeption und erste Resultate aus einer Schulleistungsstudie. *Zeitschrift für Pädagogik*, 47, 179-200.

Klieme, E., Ebach, J., Funke, J., & Reeff, J.-P. (in press). *Analytical problem solving in real world contexts. Assessment framework for the International Adult Literacy and Life Skills Survey*. Ottawa: Statistics Canada, and Washington, DC: National Center of Education Statistics.

Jürgen Baumert. Max Planck Institute for Human Development, Lentzeallee 94, D-14195 Berlin, Germany, sekbaumert@mpib-berlin.mpg.de, <http://www.mpib-berlin.mpg.de>

Current theme of research:

School effectiveness. Instructional quality. Problem solving in mathematics and science. Large-scale assessment.

Most relevant publications in the field of Psychology of Education:

- Baumert, J., & Köller, O. (2000). Unterrichtsgestaltung, verständnisvolles Lernen und multiple Zielerreichung im Mathematik- und Physikunterricht der gymnasialen Oberstufe. In J. Baumert, W. Bos, & R. Lehmann (Eds.), *TIMSS/III. Dritte Internationale Mathematik- und Naturwissenschaftsstudie – Mathematische und naturwissenschaftliche Bildung am Ende der Schullaufbahn: Vol. II. Mathematische und physikalische Kompetenzen am Ende der gymnasialen Oberstufe* (pp. 271-315). Opladen: Leske + Budrich.
- Baumert, J., & Köller, O. (2000). Motivation, Fachwahlen, selbstreguliertes Lernen und Fachleistungen im Mathematik- und Physikunterricht der gymnasialen Oberstufe. In J. Baumert, W. Bos, & R. Lehmann (Eds.), *TIMSS/III. Dritte Internationale Mathematik- und Naturwissenschaftsstudie – Mathematische und naturwissenschaftliche Bildung am Ende der Schullaufbahn: Vol. II. Mathematische und physikalische Kompetenzen am Ende der gymnasialen Oberstufe* (pp. 181-213). Opladen: Leske + Budrich.
- Köller, O., Schnabel, K. U., & Baumert, J. (in press). Does interest matter? The relationship between academic interest and achievement in mathematics. *Journal for Research in Mathematics Education*.
- Marsh, H.W., Köller, O., & Baumert, J. (2001). Reunification of East and West German school systems: Longitudinal multilevel modeling study of the big-fish-little-pond effect on academic self-concept. *American Educational Research Journal*, 38.