Chapter 11
The Diagnostic Skills of Mathematics Teachers

Martin Brunner, Yvonne Anders, Axinja Hachfeld, and Stefan Krauss

11.1 Teachers' Diagnostic Skills: Definition and Relevance

Teacher judgments of students' academic achievement provide vital information for both research and applied assessment worldwide (for an overview, see Meisels et al. 2001). It therefore comes as no surprise that teachers' diagnostic skills (an important component of their professional competence) have received considerable attention in the ongoing debate on pre- and in-service teacher training (see, e.g., Baumert and Kunter 2006, for Germany). Teachers' *diagnostic skills* can be defined as their ability (a) to accurately judge student characteristics relevant to learning and achievement and (b) to appropriately gauge the demands of learning activities and tasks (Artelt and Gräsel 2009; Schrader 1989, 2009). Ideally, teachers apply their diagnostic skills not only when devising, correcting, and grading tests and examinations but especially when preparing lessons and monitoring students' understanding during the learning process (Baumert and Kunter 2006; Hoge and Coladarci 1989; Meisels et al. 2001; National Board for Professional Teaching

M. Brunner (✉)
Chair for Evaluation and Quality Management in Education, Free University of Berlin
and Berlin-Brandenburg Institute for School Quality, Otto-von-Simson-Str. 15,
D-14195 Berlin, Germany
e-mail: martin.brunner@isq-bb.de

Y. Anders • A. Hachfeld
Chair for Early Childhood Education, Free University of Berlin,
Habelschwerdter Allee 45, D-14195 Berlin, Germany
e-mail: yvonne.anders@fu-berlin.de; axinja.hachfeld@fu-berlin.de

S. Krauss
Faculty of Mathematics, Education of Mathematics, University of Regensburg,
Universitätsstraße 31, 93053 Regensburg, Germany
e-mail: Stefan1.Krauss@mathematik.uni-regensburg.de

M. Kunter et al. (eds.), *Cognitive Activation in the Mathematics Classroom and Professional Competence of Teachers*, Mathematics Teacher Education 8,
DOI 10.1007/978-1-4614-5149-5_11, © Springer Science+Business Media New York 2013

Standards 2002; Shulman 1987). Teachers' diagnostic skills are thus of particular relevance in two respects: in the assignment of grades and for student progress.

Given the critical importance of grades for students' educational careers and life chances in general, the relevance of teachers' diagnostic skills in the context of grading is clear (Hoge and Coladarci 1989; Meisels et al. 2001; Tent 2001). Grades are decisive for promotion to the next grade level at the end of the school year, and students' allocation to different school types and tracks depends primarily on the grades they obtain. Finally, grades feed into the qualifications awarded, which in turn regulate access to many careers. It is therefore important that teacher judgments not be biased or inaccurate but that teachers demonstrate sound diagnostic skills in their grading practice (see also Dünnebier et al. 2009).

The relevance of diagnostic skills for student progress can be explained by reference to current models of instructional quality. For example, the COACTIV model of instructional quality (see Chap. 6) sees instruction as an *opportunity structure* for *insightful learning processes in schools.* From this perspective, the primary task of instruction is to facilitate students' independent and active engagement with their existing knowledge and with new instructional content. Teachers' diagnostic skills come into play in their implementation of two central dimensions of instructional quality. First, the more instruction succeeds in facilitating students' active cognitive engagement with lesson content, the higher the *potential for cognitive activation.* In particular, tasks that build on students' prior knowledge and call their existing knowledge into question are considered to be cognitively activating. In order to be able to select appropriate tasks, teachers need to be able to accurately gauge the difficulty and cognitive demands of tasks, on the one hand, and the prior knowledge of their students, on the other. Second, a supportive learning environment is needed to encourage student take-up of cognitively activating learning opportunities (Pintrich et al. 1993). In order to provide *individual learning support,* teachers must be able to notice when students are having difficulty understanding. In sum, teachers ideally use their diagnostic skills (1) to gauge the cognitive demands and difficulties of tasks and to evaluate (2) the prior knowledge and (3) comprehension problems of the students in their class. The better they succeed in doing so, the better able they are to create opportunity structures for insightful learning processes that are adapted to the abilities and needs of their students (see also Corno and Snow 1986; Helmke 2003; Hoge and Coladarci 1989; National Council of Teachers of Mathematics 2000; Shulman 1987).

In this chapter, we examine the diagnostic skills of mathematics teachers. It follows from the reasoning that these skills are relevant to student progress that mathematics teachers' diagnostic skills necessitate the integration of various facets from two of the key domains of teacher knowledge defined in the *COACTIV model of teachers' professional competence* (see also Chap. 2): pedagogical content knowledge (see Chap. 8) and pedagogical/psychological knowledge (see Chap. 10; Fig. 11.1). One important facet of (nonsubject specific) *pedagogical/psychological knowledge* concerns the assessment of student achievement (e.g., knowledge of the testing and evaluation of student achievement). Mathematics teachers need this knowledge of content and methods in order to gauge their students' learning motivation and prior knowledge in mathematics as key student characteristics relevant to

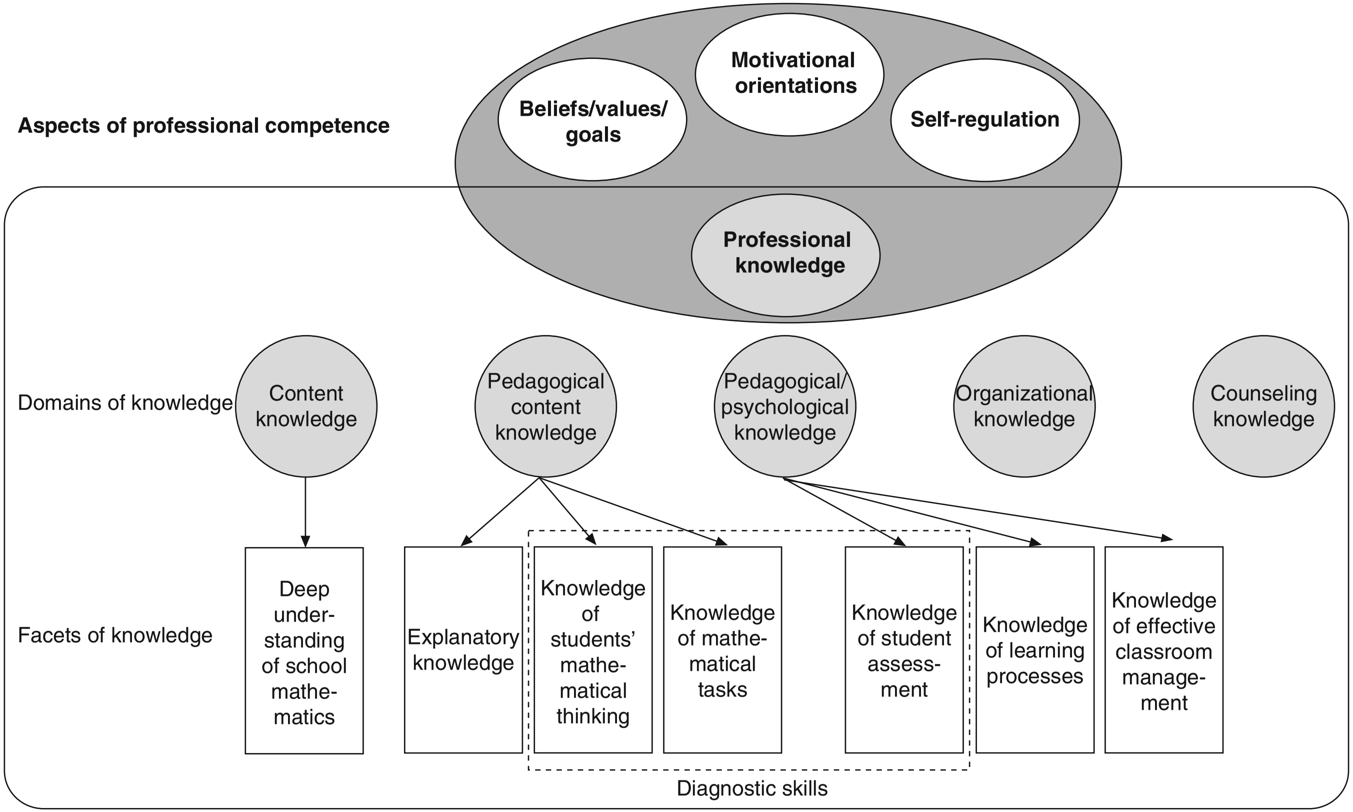

Aspects of professional competence
Beliefs/values/ goals
Motivational orientations
Self-regulation
Professional knowledge
Domains of knowledge
Content knowledge
Pedagogical content knowledge
Pedagogical/ psychological knowledge
Organizational knowledge
Counseling knowledge
Facets of knowledge
Deep under-standing of school mathe-matics
Explanatory knowledge
Knowledge of students' mathe-matical thinking
Knowledge of mathe-matical tasks
Knowledge of student assess-ment
Knowledge of learning processes
Knowledge of effective classroom manage-ment
Diagnostic skills

Fig. 11.1 Embedding of diagnostic skills in the COACTIV model of teachers' professional competence: Diagnostic skills represent a multidimensional facet of teacher competence, integrating several facets of pedagogical content knowledge and pedagogical/psychological knowledge

learning and achievement. *Pedagogical content knowledge* is the (subject specific) knowledge needed to make mathematical content “accessible” to students. Beside knowledge of subject-specific instructional strategies, it implies knowledge of the potential of mathematical tasks and of student cognitions about the subject. Teachers’ knowledge of students’ mathematics-related cognitions is of course critical in their assessment of students’ prior mathematical knowledge; it is a major regulatory factor in the diagnostic process (e.g., teachers can select tasks specifically to test whether the students in their class hold certain mathematical misconceptions). Finally, in order to gauge the demands of learning activities and tasks, mathematics teachers require knowledge of the potential and cognitive demands of mathematical tasks. In sum, in order to accurately judge student characteristics relevant to learning and achievement as well as the demands of tasks, mathematics teachers need to integrate various facets of pedagogical/psychological knowledge and pedagogical content knowledge.

Teachers’ diagnostic skills are considered so important that they are now anchored in teacher education curricula in Germany and elsewhere (see also National Board for Professional Teaching Standards 2002). In Germany, for example, the KMK (the council of Germany’s state ministers of education) introduced binding national standards for teacher education at the start of the 2005/2006 academic year. These standards specify the “diagnosis and support of individual learning processes, measurement and evaluation of student achievement” as major focuses of the teacher education curriculum (KMK 2004, p. 5, own translation). The establishment of a number of new university chairs focusing on teachers’ diagnostic skills has been a logical consequence of this development (Artelt and Gräsel 2009, p. 157).

Despite the high political and practical relevance of teachers’ diagnostic skills, there is still a considerable need for research in this area (Schrader 2009, p. 238). In Germany, research on the topic has intensified markedly in recent years (Artelt and Gräsel 2009). In this chapter, we aim to advance this area of research by reporting and discussing selected findings from the COACTIV study on the diagnostic skills of secondary-level mathematics teachers in Germany. Specifically, we address the following questions: (1) How well are mathematics teachers able to evaluate the achievement level, distribution of achievement, and motivation of their classes? (2) Do the different indicators of diagnostic skills represent a single one-dimensional construct? (3) Do teachers’ diagnostic skills influence their students’ achievement in mathematics?

11.2 The Investigation of Diagnostic Skills in the COACTIV Study

11.2.1 Design of the COACTIV Study

The COACTIV study was conceptually and technically embedded in the German extension to the 2003 cycle of the OECD’s PISA study (Kunter et al. 2007). Students in the “PISA classes” were administered achievement tests and questionnaires tapping their learning motivation and ratings of instructional quality at the end of grade 9 and grade 10. Within the COACTIV framework, their mathematics teachers were also administered questionnaires and tests (see Chap. 5 for details of the study design). Note that the description of the sample given in Chap. 5 applies in varying degrees to the data presented in the following. In some cases, data were available for only part of the sample, resulting in varying sample sizes. However, as the sampling procedure used in the PISA study resulted in relatively large numbers of participants, the samples used to address all of the present research questions can be considered representative of the corresponding populations of secondary teachers in Germany (see also Kunter et al. 2005). A description of the German school system is provided in Chap. 3.

11.2.2 Assessment of Diagnostic Skills

In order to accurately judge (a) student characteristics relevant to learning and achievement and (b) the demands of learning activities and tasks for the students in their classes, mathematics teachers need to integrate various facets of teacher knowledge: knowledge of diagnostic methods, knowledge of the potential of mathematical tasks, and knowledge of students' mathematical cognitions. As definitions of diagnostic skills vary, in COACTIV we administered several established instruments (Hoge and Coladarci 1989; McElvany et al. 2009; Schrader 1989) targeting different objects of judgment (motivation vs. student achievement; performance on a specific task versus the full mathematics test) and different levels of judgment (individual students vs. whole class). In all cases, the accuracy of teacher judgments was determined by comparing teachers' ratings with the actual outcomes of the students in their class. The closer the agreement between the teacher judgments and these objective outcomes, the more developed the diagnostic skill in question.

At the class level, teachers were asked to provide the following ratings: "Please rate the *achievement level* of your PISA class in mathematics relative to an average class of the same school type," "Please rate the *distribution of achievement* in mathematics in your PISA class relative to an average class of the same school type," and "Please rate the *motivation* of your PISA class in mathematics relative to an average class of the same school type." All responses were given on a 5-point rating scale with the options "considerably below average" (coded 1), "somewhat below average" (coded 2), "average" (coded 3), "somewhat above average" (coded 4), and "considerably above average" (coded 5). To determine the accuracy of the teachers' judgments, we then compared their responses with the actual outcomes of their PISA classes. To this end, we first calculated quintiles for achievement level, distribution of achievement, and motivation[1] separately for each school type. Each PISA class was then assigned to one of these quintiles (see Spinath 2005, for an analogous procedure): The first quintile was coded 1, the second quintile was coded 2, etc. In a second step, we computed the difference between the teachers' ratings and these

[1] The class mean score on the effort scale (see Ramm et al. 2006) of the national PISA student questionnaire was used as a class-specific indicator of motivation in mathematics. A sample item from this scale is "In mathematics I make a real effort to understand everything."

objective quintiles. In the following, the absolute value of the difference is termed the *judgment error.* A judgment error of zero indicates that the teacher rating was fully congruent with the objective outcome. The *judgment tendency,* in contrast, reflects the degree of over- or underestimation of the actual class outcomes. Positive scores indicate that a teacher tends to overestimate students' achievement; negative scores indicate that he or she tends to underestimate their achievement.[2]

To provide further indicators of diagnostic skills at the class level, teachers were asked to estimate the percentages of high- and low-achieving students in their PISA class by answering the following questions: "Relative to other classes of the same grade and school type, please estimate the percentage of students in your PISA class performing at a *high-achievement level* (in the top third)" and "Relative to other classes of the same grade and school type, please estimate the percentage of students in your PISA class performing at a *low-achievement level* (in the bottom third)." To gauge the accuracy of these judgments, we then computed the *judgment error* in terms of the absolute difference between the teachers' judgments and the actual percentage of high- versus low-achieving students in the class.

To evaluate the accuracy of their *assessment of task demands,* we asked the teachers to estimate how many of the students in their class would be able to solve each of four tasks correctly. These tasks (see Fig. 11.2) addressed important domains of mathematical content typically covered at secondary level and were administered in the German national extension to the PISA 2003 mathematics assessment. For each task, we computed the absolute difference between the teachers' estimates and the actual proportion of correct answers in the class as a measure of judgment error. The mean judgment error across the four tasks—the *task-related judgment error*—was then calculated. A task-related judgment error of zero indicates that a teacher correctly estimated the number of correct solutions in their PISA class on all four tasks.

All of the above indicators relate to the class as a whole. To examine the teachers' ability to predict the performance of individual students, we additionally asked the teachers to consider seven *individual students,* who were drawn at random from their class. First, they rated whether or not these students would be able to solve the tasks "Kite" and "Mrs. May" correctly. We determined the accuracy of these individual teacher judgments by calculating the proportion of the 14 predictions that were correct. The theoretically possible range was thus from 0 to 1, with a score of 1 indicating that all 14 of a teacher's predictions were correct.

Finally, we asked the teachers to judge how well the same seven students performed on the PISA 2003 mathematics assessment by putting them in rank order of achievement. This rank order was compared with the students' actual rank order of achievement on the PISA mathematics assessment. To provide a measure of *diagnostic sensitivity,* we then computed the rank correlation (Spearman's ρ) of the two rank orders. The higher the diagnostic sensitivity score, the better able a teacher was to predict the rank order of achievement; a score of 1 indicates a perfect prediction.

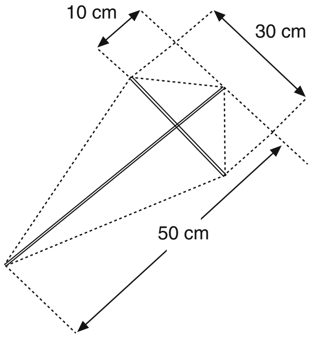
[2] Different judgment tendencies may thus result in the same judgment error scores. A teacher who overestimates the achievement level of her class by one point will have the same judgment error score as a teacher who underestimates the performance level of her class by one point.

a. "Kite"

Some students want to make kites. Peter and Rosie prepare frames out of light wooden sticks.

Then they want to stick a thin sheet of plastic film onto this frame. It has to be a single piece of film.

What is the surface area of the plastic film to be stuck on the kite?



(Drawing not to scale)

b. "Mrs May"

Mrs May runs a clothes shop. She pays a wholesale price of €150 for a dress from a supplier.

She calculates the retail price to be written on the price tag as follows: First she increases the wholesale price by 100%. Then she adds 16% tax to this new price.

What price does Mrs May write on the price tag?

c. "Sausage Stand a and b"

A class is running a sausage stand at a school fete. One student prepares a price table for bigger orders. But he makes a mistake in his calculations.

a) Put a cross in the column containing the mistake.

Number of sausages	3	4	6	8	10
Price	€3.60	€4.80	€7.20	€8.60	€12.00
	□	□	□	□	□

b) Give reasons for your decision and correct the mistake.

The mathematics teachers were asked to state which of seven students drawn at random from their class would answer the tasks "Kite" and "Mrs May" correctly. In addition, they were asked to estimate the overall percentage of students in their class who would solve each of the tasks "Kite," "Mrs May," and "Sausage Stand a and b" correctly.

Fig. 11.2 Tasks used to assess teachers' diagnostic skills

11.2.3 How Accurately Do Mathematics Teachers Judge the Achievement Level, Distribution of Achievement, and Motivation of Their Classes?

11.2.3.1 Theoretical Background

Ideally, teachers should apply their diagnostic skills to gauge the cognitive demands and difficulties of tasks, on the one hand, and to evaluate the prior knowledge and comprehension problems of the students in their class, on the other. The better they succeed in doing so, the better able they are to create opportunity structures for insightful learning processes that are adapted to the abilities and needs of their students (see Chap. 6; Corno and Snow 1986; Helmke 2003; National Council of Teachers of Mathematics 2000; Shulman 1987). These processes of adaptation may concern either individual students or the class as a whole. In order to plan effective whole-class instruction, for example, teachers need to select tasks that are appropriate to the ability and motivation of the class. Processes of adaptation at the class level thus depend on the accurate assessment of a class's achievement level, distribution of achievement, and motivation. But how accurate are the judgments of secondary-level mathematics teachers in these respects?

Previous research on teachers' diagnostic skills has focused on elementary teachers (Hoge and Coladarci 1989; Karing 2009; Schrader 1989; Spinath 2005) and primarily on individual student achievement. These studies have tended to focus on diagnostic sensitivity—that is, the accuracy of teacher judgments of rank orders of achievement. However, diagnostic sensitivity is not an appropriate measure of how accurately teachers are able to judge the achievement level or the distribution of achievement in their class—it reflects only the agreement of rank orders, irrespective of whether the absolute level and distribution of student achievement are correctly gauged.

Few studies to date have analyzed the latter two diagnostic skills, and their findings have been mixed: Some studies found that teachers tend to overestimate their students' academic functioning (Demaray and Elliot 1998; Spinath 2005); others reported very accurate judgments (see Spinath 2005, on teacher judgments of student intelligence) or underestimation of student achievement (Artelt et al. 2001; Feinberg and Shapiro 2003). Studies examining the accuracy of teacher judgments of the distribution of student outcomes within their classes have reported that the heterogeneity of both intelligence (Spinath 2005) and mathematics achievement (Schrader 1989) tend to be overestimated.

There has been little previous research on the accuracy of teacher judgments of students' motivational characteristics (Karing 2009; Spinath 2005). Hosenfeld et al. (2002) found that teachers underestimated the level of student interest in a specific lesson. Spinath (2005) found that, on average, elementary school teachers underestimated the level of their students' competence beliefs and learning motivation but overestimated their school anxiety.

In sum, previous research on the accuracy of teachers' judgments of the level and distribution of student characteristics at the class level has focused on elementary school teachers. Irrespective of the object of judgment and the particular diagnostic skill investigated, teacher judgments have relatively rarely been found to be accurate. We therefore drew on the COACTIV data to investigate whether these findings on the accuracy of teacher judgments of the level and distribution of student achievement and motivation can be generalized to mathematics teachers at lower secondary level.

11.2.3.2 Sample

The following analyses are based on data obtained from 331 mathematics teachers (42% women) who taught a grade 9 PISA class in 2003. Of these teachers, 23%

taught at a vocational-track school, 10% at a multitrack school, 26% at an intermediate-track school, 9% at a comprehensive school, and 32% at an academic-track school.

11.2.3.3 Results

In the following analyses, we focus on the accuracy of teacher judgments of their PISA class's achievement level, distribution of achievement, and motivation. How accurately did the teachers assess their class in these respects? The distribution of responses is given in Fig. 11.3. As shown, most teachers judged the achievement level, distribution of achievement, and motivation of their PISA classes to be average. Very few teachers judged their classes to be considerably above average in these respects.

How accurate were these judgments? The negative mean scores for level of achievement and motivation presented in Table 11.1 indicate that the teachers generally tended to underestimate these outcomes in their PISA classes. Teacher judgments of the distribution of achievement in their class tended to be relatively accurate. However, the high standard deviations for all three diagnostic skills indicate that teachers differed markedly in their ability to gauge these outcomes in their PISA classes.

As a further measure of the accuracy of teacher judgments, we computed Spearman rank correlations between the teacher judgments and actual class outcomes (Table 11.1). In the total sample, higher teacher judgments of achievement level ($r=0.31$), distribution of achievement ($r=0.15$), and motivation ($r=0.14$)

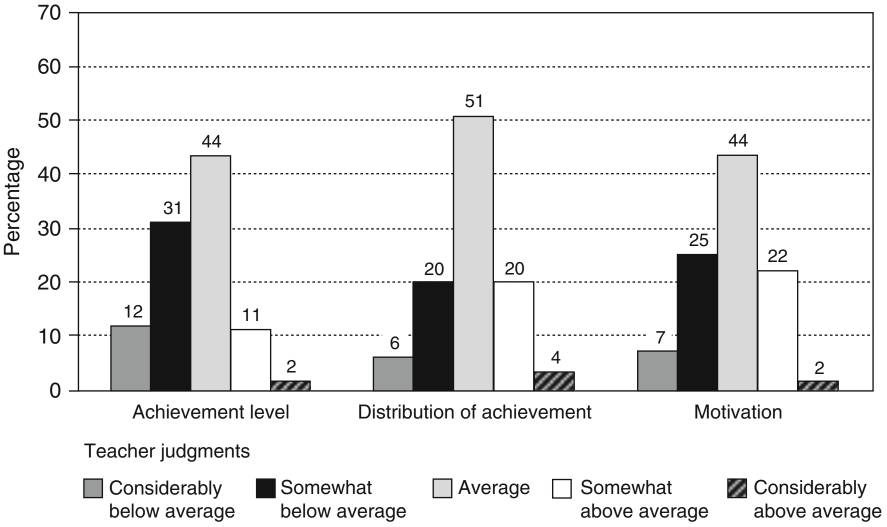


Fig. 11.3 Teacher judgments of the achievement level, range of achievement, and motivation of their PISA class in mathematics relative to an average class of the same school type. Percentage distribution of teacher responses in the full sample ($N=331$)

Table 11.1 Teacher judgments of achievement level, distribution of achievement, and motivation: descriptive statistics for judgment tendency ($N = 331$) and Spearman rank correlations between teacher judgments and the actual outcomes of their PISA class

	Judgment tendency				Correlation with class outcome		
Teacher judgments	*M*	*SD*	Min	Max	Ach. lev.	Dist.	Mot.
Achievement level	–0.43	1.43	–4	3	**0.31**	0.04	**0.11**
Distribution of achievement	–0.05	1.54	–4	4	0.03	**0.15**	0.01
Motivation	–0.24	1.58	–4	3	**0.21**	0.02	**0.14**

Note: Negative judgment tendency scores indicate that teachers underestimated the actual outcomes of the students in their class

Correlations shown in bold were statistically significant at $p < 0.05$ (two-tailed test)

Min minimum, *Max* maximum, *Ach. lev.* achievement level, *Dist.* distribution of achievement, *Mot.* motivation

were associated with higher corresponding outcomes at the class level: For example, if a teacher judged the achievement level of his or her PISA class to be above average, the mean achievement level of that class did in fact tend to be above the average for classes of the same grade level and school type. However, the weak correlations show that the overall level of accuracy was low.

This low accuracy of teacher judgments is clearly illustrated in Fig. 11.4, which sets teacher judgments in relation to actual class outcomes. For example, 49% of the teachers whose class's actual level of achievement was considerably above average (i.e., among the best 20% of PISA classes of that school type) rated their classes as just average. A similar picture emerged for the teacher judgments of distribution of achievement and motivation. Thus, very few teachers seem able to accurately assess important aspects of their class's achievement and motivation. In particular, the accuracy of teacher judgments of classes whose objective outcomes were above average was low.

11.2.4 Do the Different Indicators of Diagnostic Skills Represent a Single One-Dimensional Construct?

The previous section examined specific indicators of teachers' diagnostic skills at the class level. In this section, we shift the focus to the relations between indicators of diagnostic skills that capture different objects and levels of judgment.

11.2.4.1 Theoretical Background

Teachers' diagnostic skills can be defined as their ability (a) to accurately judge student characteristics relevant to learning and achievement and (b) to appropriately gauge the demands of learning activities and tasks (Artelt and Gräsel 2009;

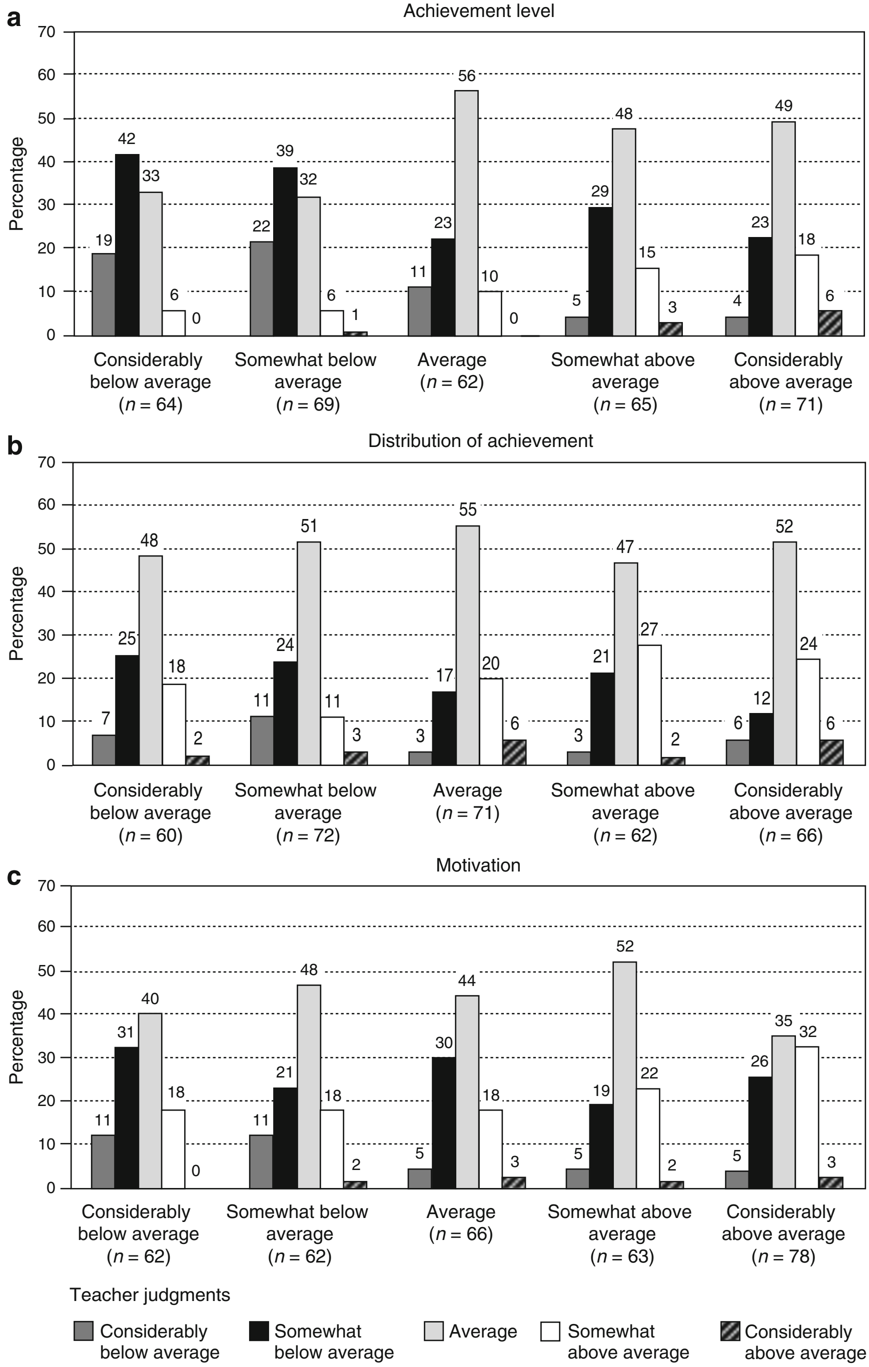


Fig. 11.4 Percentage distribution of teacher judgments of (**a**) achievement level, (**b**) distribution of achievement, and (**c**) motivation relative to actual class outcomes

Schrader 1989, 2009). This raises the question of whether (irrespective of conceptual differences in definitions of diagnostic skills; see also section "Teachers' Diagnostic Skills: Definition and Relevance") different indicators of diagnostic skills represent a single one-dimensional construct. If this were the case, it would imply that (a) indicators of diagnostic skills that capture different objects and levels of teacher judgment would intercorrelate substantially and (b) that these intercorrelations would be explained by *a single* common factor (McDonald 1981).

The dimensionality of diagnostic skills has attracted little research attention to date, and here too, the few studies conducted have focused on elementary school teachers. However, findings have been consistent across studies, with weak or no correlations being found between different indicators of diagnostic skills—this pattern of results was reported by both Schrader (1989) and Spinath (2005). The available findings thus indicate that diagnostic skills are a multidimensional construct. In this section, we examine whether this finding can be generalized to mathematics teachers at secondary level.

11.2.4.2 Sample

The following analyses are based on data obtained from 217 mathematics teachers (40% women) who taught a grade 9 PISA class in 2003 *and* for whom complete data were available on all diagnostic skills (see section "Assessment of Diagnostic Skills"). Of these teachers, 15% taught at a vocational-track school, 9% at a multi-track school, 28% at an intermediate-track school, 8% at a comprehensive school, and 40% at an academic-track school.

11.2.4.3 Results

Before we consider in detail the intercorrelations of the indicators of diagnostic skills, it is worth highlighting a descriptive finding from Table 11.2. As shown in the penultimate line of the table, the accuracy of three quarters of the teachers' predictions of whether specific students would be able to answer the "Kite" and "Mrs. May" tasks correctly did not exceed 58%. In other words, the accuracy of three quarters of the teachers' predictions was little higher than that of random guessing. One reason for this outcome is that most teachers overestimated the percentage of students in their class who would solve the two tasks correctly. The low accuracy of their predictions of individual student performance thus seems to be a logical consequence of teachers misestimating the base rate of correct solutions in the class as a whole.

We now return to the main question of this section: Do the different indicators of diagnostic skills represent a single one-dimensional construct? As Table 11.2 shows, the intercorrelations between the various indicators of diagnostic skills were weak (median $r = -0.01$; mean $r = 0.00$). Moreover, the pattern of correlations was relatively mixed (standard deviation of the correlations = 0.12). The lowest correlation

Table 11.2 Descriptive statistics and intercorrelations of the indicators of diagnostic skills ($N = 217$)

Indicators of diagnostic skills	1	2	3	4	5	6	7	8
Relating to the class as a whole								
1. JE achievement level	—							
2. JE distribution of achievement	0.11	—						
3. JE % students in bottom third of achievement distribution	**0.27**	0.02	—					
4. JE % students in top third of achievement distribution	**0.33**	–0.01	–0.07	—				
5. JE motivation	–0.11	–0.04	0.03	–0.07	—			
Relating to mathematics tasks and the class as a whole								
6. Task-related JE	–0.05	0.11	0.09	–0.01	0.01	—		
Relating to individual students								
7. Accuracy of prediction of ability to solve mathematics tasks	–0.01	–0.06	0.06	–0.12	–0.04	**–0.34**	—	
8. Diagnostic sensitivity	–0.04	0.05	–0.07	–0.02	0.05	–0.06	0.12	—
Descriptive statistics								
M	1.18	1.22	0.15	0.19	1.29	0.27	0.51	0.39
SD	0.95	0.94	0.14	0.14	0.90	0.11	0.15	0.36
Minimum	0.00	0.00	0.00	0.00	0.00	0.07	0.14	–0.71
25th percentile	0.00	1.00	0.04	0.08	1.00	0.18	0.43	0.16
Median	1.00	1.00	0.10	0.17	1.00	0.26	0.50	0.43
75th percentile	2.00	2.00	0.22	0.27	2.00	0.35	0.58	0.69
Maximum	4.00	4.00	0.80	0.69	4.00	0.56	0.93	0.94

Note: Correlations shown in bold were statistically significant at $p < 0.05$ (two-tailed test).
JE judgment error

coefficient ($r = –0.34$) was between the task-related judgment error and the accuracy of teachers' predictions of whether specific students would be able to solve the "Kite" and "Mrs. May" tasks correctly. This finding again indicates that the accuracy of teachers' judgments of individual students' performance decreased as a function of their misestimation of the base rate of correct solutions in the class as a whole. The highest correlation coefficient ($r = 0.33$) was between the error in teachers' judgments of the class achievement level and the error in their judgments of the percentage of students in their class performing in the top third of the achievement distribution relative to other classes of the same grade and school type. The correlation with the error in teacher judgments of the percentage of students performing in the bottom third of the achievement distribution was of a similar magnitude. These relatively high correlations can be attributed to two main sources. First, teachers' judgments of the mean achievement level of their PISA class are doubtless affected by their estimates of the proportion of high- versus low-achieving students in their class. Second, the actual proportion of students in the top (or bottom) third of the achievement distribution strongly influences the actual achievement level of the whole class. Given that both teacher judgments and the actual proportion of students in the top (bottom) third of the achievement distribution or the actual class mean feed into these indicators of teachers' diagnostic skills, the relatively high correlations are not surprising (see also Cohen et al. 2003).

In view of the generally weak intercorrelations of the different indicators of diagnostic skills, we did not conduct further factor analyses—it can be assumed a priori that a one-factor model cannot explain this pattern of intercorrelations. In conclusion, our analyses indicate that the different indicators of mathematics teachers' diagnostic skills at secondary level do not represent a one-dimensional but a multidimensional construct.

11.2.5 Do Teachers' Diagnostic Skills Influence Students' Mathematics Achievement?

11.2.5.1 Theoretical Background

According to current thinking in instructional research, teachers' diagnostic skills are highly relevant for the progress of the students in their classes (see also section "Teachers' Diagnostic Skills: Definition and Relevance"). Two mechanisms are thought to underlie the assumed positive effects. First, teachers with good diagnostic skills are able to accurately assess student characteristics relevant to learning and achievement on both the individual and the class level. Second, they are able to judge the difficulty of instructional material and its potential for cognitive activation (Anders et al. 2010). These evaluations, and the associated processes of adaptation, are expected to result in teachers providing individual learning support for their students, on the one hand, and developing the potential for cognitive activation in their lessons, on the other. In so doing, teachers create opportunity structures for insightful learning processes.

Although this reasoning seems plausible, the empirical data to support it are both limited and inconclusive, as the findings of previous studies have been mixed. Fisher et al. (1978) found a positive relationship between teachers' ability to judge the difficulty of the tasks in a mathematics test and their students' achievement and engagement in the subject. Lehmann et al. (2000) examined the relationship between teachers' ability to gauge the difficulty of individual mathematics tasks for the students in one of their classes and those students' test scores at the end of the school year. Their findings were mixed, with positive relations emerging for some school types and grades but not for others. Findings reported by Helmke and Schrader indicated that teachers' instructional practice mediates the relationship between high diagnostic skills and student achievement gains in mathematics (Helmke and Schrader 1987; Schrader 1989): The greatest learning gains were observed in classes in which teacher judgments were accurate and instructional quality was high.

In sum, more empirical research is needed into the effects of teachers' diagnostic skills on student progress, especially as the results of previous studies have been mixed. In this section, we therefore examine the extent to which mathematics teachers' diagnostic skills were positively related to student outcomes when relevant student baseline variables are controlled. As a detailed description of all COACTIV findings on this research question is available in Anders et al. (2010), the following account is limited to the central findings.

11.2.5.2 Results

The following analyses are based on data obtained from 155 mathematics teachers (47% women) and from 3,483 students in the PISA classes. In view of our finding (see section "Do the Different Indicators of Diagnostic Skills Represent a Single One-Dimensional Construct?") that diagnostic skills are a multidimensional construct, the following analyses focus on two central indicators: task-related judgment error for the class as a whole (in terms of the mean judgment error on the items "Sausage Stand a and b") and diagnostic sensitivity. The central dependent variable in these analyses was grade 10 mathematics achievement. Because (in contrast to randomized experiments) students are not assigned to classes or school types at random, we used hierarchical linear modeling (HLM; Raudenbush and Bryk 2002) to control for a number of variables at the student and class levels, thus isolating the potential effect of diagnostic skills on mathematics achievement. The control variables at student level were selected to model the process of allocation to the different types of secondary school (see Baumert et al. 2010). This process depends strongly on the tracking recommendation made by the elementary teacher, which is based largely on the student's mathematical literacy, reading literacy, and (basic) cognitive abilities. At the same time, family background (parental education and occupation; immigration status) is also an important determinant of tracking decisions. At the class level, we controlled for several important context variables and teacher characteristics that are thought to positively affect the achievement of the students in a class. These include task potential as an indicator of the potential for cognitive activation in lessons, class size, and the teacher's career and teaching experience.

The major findings of the HLM analyses were that both indicators of diagnostic skills were statistically significantly associated with students' mathematics achievement (see Tymms 2004, for the computation of the ES_{HLM} effect size): The smaller a mathematics teacher's task-related judgment error, the higher the mathematics achievement of his or her students in grade 10 ($ES_{HLM}=-0.14$). Higher diagnostic sensitivity was also associated with higher mathematics achievement in grade 10 ($ES_{HLM}=0.16$). When student background variables and context conditions at class level were controlled, those classes whose teachers gave more accurate judgments of (1) task-related difficulty and (2) the rank order of the students in their class achieved higher scores on the grade 10 mathematics assessment. Given that the achievement gain in mathematics from grade 9 to grade 10 was around 0.3 standard deviations, the seemingly "small" effect sizes of the indicators of diagnostic skills, with absolute values of around 0.15 standard deviations, are clearly of practical relevance (Baumert and Artelt 2002; Hill et al. 2008).

11.3 Discussion

11.3.1 Summary

To create opportunity structures for insightful learning processes, teachers need to adapt their instruction to the abilities and needs of their students (see Chap. 6; Corno and Snow 1986; Helmke 2003; Hoge and Coladarci 1989; National Council of Teachers of Mathematics 2000; Shulman 1987). Diagnostic skills play an important role in this context. At the same time, sound diagnostic skills are crucial in grading process (Dünnebier et al. 2009; Meisels et al. 2001). In this chapter, we reported selected findings from COACTIV on the diagnostic skills of secondary-level mathematics teachers in Germany. First, we presented the instruments used, which targeted different objects of judgment (motivation vs. student achievement; performance on a specific task vs. the full mathematics test) and different levels of judgment (individual students vs. whole classes). Our analyses were based on data obtained from a large heterogeneous sample of lower secondary mathematics teachers who participated in the COACTIV study. Our responses to the three research questions can be summarized as follows: (1) The accuracy of teachers' judgments of their classes' achievement level, distribution of achievement, and motivation is relatively low. (2) Diagnostic skills do not represent a one-dimensional but a multidimensional construct. (3) Teachers' diagnostic skills (in terms of tasks-related judgment error and diagnostic sensitivity) have a positive influence on their students' achievement gains in mathematics.

11.3.2 Strengths and Limitations of the Investigation of Diagnostic Skills in COACTIV

Because the COACTIV study was embedded within the longitudinal PISA study, we were able to (1) investigate the diagnostic skills of a large and (roughly) representative sample of lower secondary mathematics teachers in Germany and (2) examine the effects of teachers' diagnostic skills on their students' mathematics achievement over time. Previous studies of diagnostic skills have focused on elementary teachers. The results of the present study allowed many of these previous findings to be generalized to secondary teachers. The question of generalizability was by no means trivial, as elementary and secondary teachings differ in numerous respects that might influence the accuracy of teacher judgments (e.g., elementary school teachers tend to teach the same class several subjects, whereas secondary school teachers tend to teach the same subject(s) to several classes; teacher education differs; the ability mix of classes differs; for a summary, see Karing 2009).

Despite the strengths of the COACTIV study, some of the findings reported in this chapter require qualification. Our findings on mathematics teachers' diagnostic skills are based on selected indicators that have previously been administered in the same form as in other studies (Hoge and Coladarci 1989; Lorenz and Artelt 2009; Schrader 1989; Spinath 2005). However, these indicators cover only certain aspects of the diagnostic process in schools (Artelt and Gräsel 2009). In order to gain a thorough understanding of the role of diagnostic skills in instruction, it would be necessary to assess not only various indicators of teachers' judgment accuracy but also, for example, their knowledge of different methods of assessment, knowledge of the effects of different reference norms, knowledge of typical student errors, and knowledge of the diagnostic potential of tasks. This combination of the various declarative and procedural knowledge components feeding into the diagnostic process can be summarized and analyzed under the broader construct of what Helmke (2003) has termed *diagnostic expertise.*

All of the findings reported here relate either to a whole class or to individual students in a class. We did *not* examine whether the accuracy of teacher judgments depends on characteristics of the class, the students, or the tasks evaluated (see also Hoge and Coladarci 1989). However, preliminary findings based on the COACTIV data point to a complex interaction of student and task characteristics. For example, the accuracy of teacher predictions of student performance on linguistically complex tasks is lower for students with German as a second language than for students whose first language is German (Hachfeld et al. 2010).

It is also important to bear two points in mind when considering the reported accuracy of teacher judgments. First, some studies have shown that the accuracy of teacher judgments is affected by the objective of the assessment: Accuracy tends to be higher in high-stakes contexts (Chen and Chaiken 1999; Krolak-Schwerdt et al. 2009). In COACTIV, the teacher judgments had no consequences for either the teachers or the students assessed (see also Lorenz and Artelt 2009). Second, it would have been very difficult for teachers to judge the student outcomes under investigation in their PISA class. Prior to the COACTIV study, most of the participating mathematics teachers had not received any feedback from standardized national assessments on the performance or motivation of their students. Both of these factors offer an explanation of why the level of diagnostic sensitivity in our study (median: $\rho = 0.43$) was below that reported by Hoge and Coladarci (1989) in their meta-analysis, where diagnostic sensitivity scores ranged between $r = 0.48$ and $r = 0.92$, with a median of $r = 0.69$. The low accuracy of teacher judgments in the COACTIV sample is therefore not surprising, and the results reported in this chapter can be assumed to reflect the lower rather than the upper boundary of mathematics teachers' judgment accuracy.

11.3.3 Implications

These findings highlight the great potential of the national assessments of student achievement (Helmke et al. 2004; Lorenz and Artelt 2009) that are now being carried out in many countries (e.g., Germany, Luxembourg, and Austria). These assessments can inform teachers about their students' absolute achievement level (e.g., in terms of proficiency levels) and relative achievement level (e.g., compared with the means of other classes) or how many students in their class are able to solve specific tasks correctly. Depending on the applicable data protection regulations, it may also be possible to provide feedback on individual students' achievement. This kind of feedback, in combination with a greater focus on diagnostic skills in pre- and in-service teacher training, can certainly help to enhance the accuracy of teacher judgments. As the findings of the present study (see section "Do Teachers' Diagnostic Skills Influence Students' Mathematics Achievement?") show, this kind of approach has the potential to both increase instructional effectiveness and foster (greater) consistency in grading standards. Although all students with the same level of achievement should theoretically be awarded the same grades, this is currently not the case (at least) in Germany (Baumert et al. 2003). Given the far-reaching implications that grades have for students' careers and lives, calls for measures to improve teachers' diagnostic skills thus seem entirely justified (Dünnebier et al. 2009; Spinath 2005).

References

Anders Y, Kunter M, Brunner M, Krauss S, Baumert J (2010) Diagnostische Fähigkeiten von Mathematiklehrkräften und ihre Auswirkungen auf die Leistungen ihrer Schülerinnen und Schüler [Mathematics teachers' diagnostic skills and their impact on students' achievements]. Psychologie in Erziehung und Unterricht 3:175–192. doi:10.2378/peu2010.art13d

Artelt C, Gräsel C (2009) Diagnostische Kompetenz von Lehrkräften. Gasteditorial [Diagnostic competence of teachers. Guest editorial]. Zeitschrift für Pädagogische Psychologie 23: 157–160. doi:10.1024/1010-0652.23.34.157

Artelt C, Stanat P, Schneider W, Schiefele U (2001) Lesekompetenz: Testkonzeption und Ergebnisse [Reading literacy: test design and results]. In: Baumert J, Klieme E, Neubrand M, Prenzel M, Schiefele U, Schneider W, … Weiß M (eds) PISA 2000: Basiskompetenzen von Schülerinnen und Schülern im internationalen Vergleich. Leske + Budrich, Opladen, pp 69–137

Baumert J, Artelt C (2002) Bereichsübergreifende Perspektiven [Cross-domain perspectives]. In: Baumert J, Artelt C, Klieme E, Neubrand M, Prenzel M, Schiefele U, … Weiß M (eds) PISA 2000: Die Länder der Bundesrepublik Deutschland im Vergleich. Leske + Budrich, Opladen, pp 219–236

Baumert J, Kunter M (2006) Stichwort: Professionelle Kompetenz von Lehrkräften [Teachers' professional competence]. Zeitschrift für Erziehungswissenschaft 9:469–520. doi:10.1007/s11618-006-0165-2

Baumert J, Trautwein U, Artelt C (2003) Schulumwelten – institutionelle Bedingungen des Lehrens und Lernens [School environments: institutional conditions of teaching and learning]. In: Baumert J, Artelt C, Klieme E, Neubrand M, Prenzel M, Schiefele U, … Weiß M (eds) PISA 2000: Ein differenzierter Blick auf die Länder der Bundesrepublik Deutschland. Leske + Budrich, Opladen, pp 261–332

Baumert J, Kunter M, Blum W, Brunner M, Voss T, Jordan A, … Tsai Y-M (2010) Teachers' mathematical knowledge, cognitive activation in the classroom, and student progress. Am Educ Res J 47:133–180. doi:10.3102/0002831209345157

Chen S, Chaiken S (1999) The heuristic-systematic model in its broader context. In: Chaiken S, Trope Y (eds) Dual process theories in social psychology. Guilford, New York, pp 73–96

Cohen J, Cohen P, West SG, Aiken LS (2003) Applied multiple regression/correlation analysis for the behavioral sciences, 3rd edn. Erlbaum, Mahwah

Corno L, Snow RE (1986) Adapting teaching to individual differences among learners. In: Wittrock MC (ed) Handbook of research on teaching, 3rd edn. Macmillan, New York, pp 605–629

Demaray MK, Elliot SN (1998) Teachers' judgments of students' academic functioning: a comparison of actual and predicted performances. School Psychol Q 13:8–24. doi:10.1037/h0088969

Dünnebier K, Gräsel C, Krolak-Schwerdt S (2009) Urteilsverzerrungen in der schulischen Leistungsbeurteilung: Eine experimentelle Studie zu Ankereffekten [Biases in teachers' assessments of student performance: an experimental study of anchoring effects]. Zeitschrift für Pädagogische Psychologie 23:187–195. doi:10.1024/1010-0652.23.34.187

Feinberg AB, Shapiro ES (2003) Accuracy of teacher judgements in predicting oral reading fluency. School Psychol Q 18:52–65. doi:10.1521/scpq.18.1.52.20876

Fisher CW, Filby N, Marliave R, Cahen LS, Dishaw MM, Moore J (1978) Teaching behaviors, academic learning time, and student achievement: final report of phase III-B, Beginning Teacher Evaluation Study. Far West Laboratory, San Francisco

Hachfeld A, Anders Y, Schroeder S, Stanat P, Kunter M (2010) Does immigration background matter? How teachers' predictions of students' performance relate to student background. Int J Educ Res 49:78–91. doi:10.1016/j.ijer.2010.09.002

Helmke A (2003) Unterrichtsqualität erfassen, bewerten, verbessern [Measuring, evaluating, and improving instructional quality]. Kallmeyersche Verlagsbuchhandlung, Seelze

Helmke A, Schrader F-W (1987) Interactional effects of instructional quality and teacher judgement accuracy on achievement. Teach Teach Educ 3(2):91–98. doi:10.1016/0742-051X(87)90010-2

Helmke A, Hosenfeld I, Schrader F-W (2004) Vergleichsarbeiten als Instrument zur Verbesserung der Diagnosekompetenz von Lehrkräften [State-wide assessments as an instrument to improve teachers' diagnostic skills]. In: Arnold R, Griese C (eds) Schulleitung und Schulentwicklung. Schneider Verlag Hohengehren, Baltmannsweiler, pp 119–144

Hill CJ, Bloom HS, Black AR, Lipsey MW (2008) Empirical benchmarks for interpreting effect sizes in research. Child Dev Perspect 2:172–177. doi:10.1111/j.1750-8606.2008.00061.x

Hoge RD, Coladarci T (1989) Teacher-based judgments of academic achievement: a review of the literature. Rev Educ Res 59(3):297–313. doi:10.3102/00346543059003297

Hosenfeld I, Helmke A, Schrader F-W (2002) Diagnostische Kompetenz: Unterrichts- und lernrelevante Schülermerkmale und deren Einschätzung durch Lehrkräfte in der Unterrichtsstudie SALVE [Diagnostic skills: student characteristics relevant to teaching and learning and their evaluation by teachers in the SALVE study]. Zeitschrift für Pädagogik 45 Beiheft:65–82

Karing C (2009) Diagnostische Kompetenz von Grundschul- und Gymnasiallehrkräften im Leistungsbereich und im Bereich Interessen [Diagnostic competence of elementary and secondary school teachers in the domains of competence and interests]. Zeitschrift für Pädagogische Psychologie 23:197–209. doi:10.1024/1010-0652.23.34.197

KMK – Secretariat of the Standing Conference of the Ministers of Education and Cultural Affairs of the Länder in the Federal Republic of Germany (2004) Standards für die Lehrerbildung: Bildungswissenschaften. Beschluss der Kultusministerkonferenz vom 16.12.2004 [Standards for teacher training: educational science. KMK resolution of 16.12.2004]. Retrieved from http://www.kmk.org/fileadmin/veroeffentlichungen_beschluesse/2004/2004_12_16-Standards-Lehrerbildung.pdf

Krolak-Schwerdt S, Böhmer M, Gräsel C (2009) Verarbeitung von schülerbezogener Information als zielgeleiteter Prozess. Der Lehrer als „flexibler Denker" [Goal-directed processing of students' attributes: the teacher as "flexible thinker"]. Zeitschrift für Pädagogische Psychologie 23:175–186. doi:10.1024/1010-0652.23.34.175

Kunter M, Brunner M, Baumert J, Klusmann U, Krauss S, Blum W, … Neubrand M (2005) Der Mathematikunterricht der PISA-Schülerinnen und -Schüler: Schulformunterschiede in der Unterrichtsqualität [Quality of mathematics instruction across school types: findings from PISA 2003]. Zeitschrift für Erziehungswissenschaft 8:502–520. doi:10.1007/s11618-005-0156-8

Kunter M, Klusmann U, Dubberke T, Baumert J, Blum W, Brunner M, … Tsai Y-M (2007) Linking aspects of teacher competence to their instruction: results from the COACTIV project. In: Prenzel M (ed) Studies on the educational quality of schools: the final report on the DFG Priority Programme. Waxmann, Münster, pp 39–59

Lehmann RH, Peek R, Gänsfuß R, Lutkat S, Mücke S, Barth I (2000) Qualitätsuntersuchungen an Schulen zum Unterricht in Mathematik (QuaSUM) [Evaluations of instructional quality in mathematics (QuaSUM)]. Ministerium für Bildung, Jugend und Sport des Landes Brandenburg (MBJS), Potsdam

Lorenz C, Artelt C (2009) Fachspezifität und Stabilität diagnostischer Kompetenz von Grundschullehrkräften in den Fächern Deutsch und Mathematik [Domain specificity and stability of diagnostic competence among primary school teachers in the school subjects of German and mathematics]. Zeitschrift für Pädagogische Psychologie 23:211–222. doi:10.1024/1010-0652.23.34.211

McDonald RP (1981) The dimensionality of tests and items. Br J Math Stat Psychol 34:100–117

McElvany N, Schroeder S, Hachfeld A, Baumert J, Richter T, Schnotz W, … Ullrich M (2009) Diagnostische Fähigkeiten von Lehrkräften bei der Einschätzung von Schülerleistungen und Aufgabenschwierigkeiten bei Lernmedien mit instruktionalen Bildern [Teachers' diagnostic skills to judge student performance and task difficulty when learning materials include instructional pictures]. Zeitschrift für Pädagogische Psychologie 23:223–235. doi:10.1024/1010-0652.23.34.223

Meisels SJ, Bickel DD, Nicholson J, Xue Y, Atkins-Burnett S (2001) Trusting teachers' judgments: a validity study of a curriculum-embedded performance assessment in kindergarten to grade 3. Am Educ Res J 38:73–95. doi:10.3102/00028312038001073

National Board for Professional Teaching Standards (2002) What teachers should know and be able to do. National Board for Professional Teaching Standards, Arlington

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

Pintrich PR, Marx RW, Boyle RA (1993) Beyond cold conceptual change: the role of motivational beliefs and classroom contextual factors in the process of conceptual change. Rev Educ Res 63:167–199

Ramm G, Prenzel M, Baumert J, Blum W, Lehmann R, Leutner D, … Schiefele U (eds) (2006) PISA 2003: Dokumentation der Erhebungsinstrumente [PISA 2003: documentation of study instruments]. Waxmann, Münster

Raudenbush SW, Bryk AS (2002) Hierarchical linear models, 2nd edn. Sage, Thousand Oaks

Schrader F-W (1989) Diagnostische Kompetenzen von Lehrern und ihre Bedeutung für die Gestaltung und Effektivität des Unterrichts [Teachers' diagnostic skills and their meaning for the delivery and effectiveness of instruction]. Lang, Frankfurt am Main

Schrader F-W (2009) Anmerkungen zum Themenschwerpunkt Diagnostische Kompetenz von Lehrkräften [The diagnostic competence of teachers]. Zeitschrift für Pädagogische Psychologie 23:237–245. doi:10.1024/1010-0652.23.34.237

Shulman LS (1987) Knowledge and teaching: foundations of the new reform. Harv Educ Rev 57:1–22

Spinath B (2005) Akkuratheit der Einschätzung von Schülermerkmalen durch Lehrer und das Konstrukt der diagnostischen Kompetenz [Accuracy of teacher judgments of student characteristics and the construct of diagnostic competence]. Zeitschrift für Pädagogische Psychologie 19:85–95. doi:10.1024/1010-0652.19.12.85

Tent L (2001) Zensuren [Grades]. In: Rost DH (ed) Handwörterbuch Pädagogische Psychologie, 2nd edn. Belz PVU, Weinheim, pp 805–811

Tymms P (2004) Effect sizes in multilevel models. In: Schagen I, Elliot K (eds) But what does it mean? The use of effect sizes in educational research. National Foundation for Educational Research, London, pp 55–66