Diagnostic Skills of Mathematics Teachers in the COACTIV Study

**Karin Binder, Stefan Krauss, Sven Hilbert, Martin Brunner,
Yvonne Anders, and Mareike Kunter**

In the present chapter we introduce theoretical and empirical approaches to the construct of diagnostic competence within the COACTIV research program. We report eight conceptualizations and operationalizations of diagnostics skills and add in addition three constructs that seem to be close to diagnostic skills. Correlational analyses reveal only moderate and unsystematic relationships. However, our analyses showed the expected differences between school types. The chapter concludes with structural equation models in which the predictive validity of diagnostic skills for mathematical achievement of students is analyzed (both with black box and with mediation models).

1 Introduction

Diagnostic skills of teachers are – among other competence aspects such as professional knowledge, certain beliefs, motivational orientation or self-regulation – considered to be relevant both for planning lessons and for teaching. In the COACTIV study (Cognitive Activation in the Mathematics Classroom and Professional Competence of Teachers; Kunter, Baumert et al., 2013) various facets of these competence aspects of mathematics teachers were assessed including several facets of

K. Binder (✉) • S. Krauss • S. Hilbert
University of Regensburg, Regensburg, Germany
e-mail: Karin.Binder@mathematik.uni-regensburg.de

M. Brunner
University of Potsdam, Potsdam, Germany

Y. Anders
Free University of Berlin, Berlin, Germany

M. Kunter
Goethe University Frankfurt, Frankfurt, Germany

T. Leuders et al. (eds.), *Diagnostic Competence of Mathematics Teachers*,
Mathematics Teacher Education 11, DOI 10.1007/978-3-319-66327-2_2

diagnostic skills. In this chapter, we review the COACTIV results with respect to diagnostic competence published so far (e.g., Anders, Kunter, Brunner, Krauss, & Baumert, 2010; Brunner, Anders, Hachfeld, & Krauss, 2013; Krauss & Brunner, 2011) and add new analyses on: (1) relationships between the different aspects of diagnostic skills, (2) respective school type differences and (3) the predictive validity for teachers' lesson quality and student learning.

As the theoretical framework on diagnostic competence is introduced in several chapters of the present publication (e.g., Leuders, T., Dörfler, Leuders, J., & Philipp, 2018; see also Südkamp & Praetorius, 2017), we will – after a short introduction into the COACTIV research program as a whole – concentrate on conceptualizations and operationalization of the related constructs in the COACTIV study.

2 The COACTIV Framework

The German COACTIV 2003/2004 research program (Cognitive Activation in the Mathematics Classroom and Professional Competence of Teachers) empirically examined a large representative sample of secondary mathematics teachers whose classes participated in the German PISA study and its longitudinal extension during 2003/04 (for an overview, see the COACTIV compendium by Kunter, Baumert et al., 2013; for PISA 2003/2004, see Prenzel et al., 2004). The structural combination of the two large scale studies PISA and COACTIV offered a unique opportunity to collect a broad range of data about students and their teachers, and to address the connection of teacher characteristics with their lesson quality and with their students' achievement (e.g., Kunter, Klusmann et al., 2013, also see Fig. 6).

In the teacher competence model of COACTIV (Fig. 1) the overarching competence aspects are: *professional knowledge, professional beliefs, motivational orientations* and *professional self-regulation skills* (the model is explicated in detail in Baumert & Kunter, 2013). In order to empirically examine research questions regarding these competence aspects (e.g., concerning their structure, school-type differences, or their impact on student learning), one needs valid and reliable measurement instruments. In COACTIV, a variety of such instruments were developed and implemented with the sample of the "COACTIV-teachers", who taught the grade 9/10 students assessed by the PISA study in 2003/2004.

In Fig. 1 (taken from Brunner et al., 2013), diagnostic skills do not appear as an autonomous competence or knowledge domain. Instead they were allocated at the intersection of pedagogical content knowledge (PCK) and pedagogical knowledge (PK). This allocation, however, remained theoretical in nature because to date no correlations of specific diagnostic skills with teacher's PCK or aspects of PK within the COACTIV data have been reported.

When analyzing such relationships in the following, we will, in addition to diagnostic skills concerning *cognitive* dimensions (such as judging student abilities or task difficulties), also include a teacher scale of diagnostic skills with respect to *social* issues ("DSS") in the following. Furthermore, in the present chapter we will

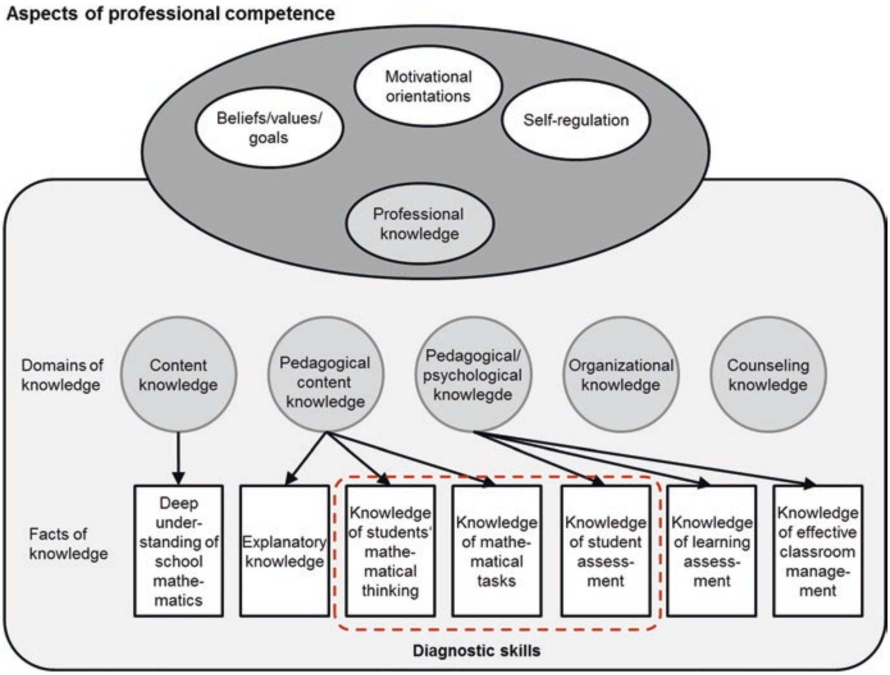

Aspects of professional competence
Beliefs/values/ goals
Motivational orientations
Self-regulation
Professional knowledge
Domains of knowledge
Content knowledge
Pedagogical content knowledge
Pedagogical/ psychological knowlegde
Organizational knowledge
Counseling knowledge
Facts of knowledge
Deep under-standing 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 assess-ment
Knowledge of effective classroom manage-ment
Diagnostic skills

Fig. 1 The COACTIV teacher competence model (Brunner et al., 2013) and the theoretical allocation of diagnostic skills (*dashed line*)

introduce another competence aspect of mathematics teachers, which is close to diagnostic skills, namely the ability to quickly judge student answers as mathematically correct or false – which we call *quick judgment skills* (QJS)[1]. It should be noted that because previous analyses based on the COACTIV data yielded only moderate correlations between different aspects of diagnostic competence (Anders et al., 2010; Brunner et al. 2013), the COACTIV research group decided to preferably use the term *diagnostic skills*.

2.1 Diagnostic Skills Assessed in COACTIV

In contrast to the domains of pedagogical content knowledge (PCK) and content knowledge (CK), both of which in COACTIV were measured by means of open test items (Krauss et al., 2013), the eight diagnostic skills, which we introduce in Sect. 2.1.1, were assessed by relating a teacher judgment on student achievement (or motivation) to the actual PISA data of his/her students. In the model of Leuders

[1] To date, this competence was described only in a German publication and not in the framework of diagnostic skills (Krauss & Brunner, 2011).

et al. (2018), all these skills (except PCK and CK) pertain to “diagnostic thinking” (Fig. 5), because they all involve perception and interpretation of the mathematical competence of both the own class as a whole and of individual students.

In Sect. 2.1.2 we describe the conceptualization and operationalization of teacher characteristics that are related to diagnostic skills such as the skill to quickly judge student answers as correct or false (QJS), PCK and DSS. The teacher competences described in Sect. 2.1.2 were assessed by paper and pencil questionnaires or tests, without the need to relate the teachers’ responses to their classes. An overview on all constructs analyzed is provided in Table 1a.

Table 1a Aspects of diagnostic skills investigated in COACTIV

Construct			
D1–D8 judgment of …		Scale	Reference
D1	Achievement level in PISA test (class average) compared to school type specific German average	1–5[a]	e.g., Brunner et al. (2013)
D2	Distribution of achievement (in class)	1–5[a]	e.g., Brunner et al. (2013)
D3	% of own students in bottom third of German achievement distribution (in class)	0–100%	e.g., Brunner et al. (2013)
D4	% of own students in top third of German achievement distribution (in class)	0–100%	e.g., Brunner et al. (2013)
D5	Motivational level (class average) compared to school type specific average	1–5[a]	e.g., Brunner et al. (2013)
D6	% correct solutions with respect to four specific PISA tasks (in class)	0–100% (for each task)	e.g., Anders et al. (2010), Brunner et al. (2013) (see Fig. 2)
D7	Solutions of two specific tasks (Kite and Mrs. May) with respect to seven specific students	2 × 7: Yes/no	e.g., Brunner et al. (2013) (see Figs. 2 and 3 left)
D8	Rank order of achievement of seven specific students in PISA (diagnostic sensitivity)	Distribution of position numbers 1–7	e.g., Anders et al. (2010), Brunner et al. (2013) (see Fig. 3 right)
QJS	Quickly classifying student answers as correct or false (12 tasks provided with respective student responses)	# correct judgments divided by the mean of 12 reaction times	e.g., Krauss and Brunner (2011) (see Fig. 4)
PCK	Pedagogical content knowledge	22 open test items	e.g., Krauss et al. (2008, 2013)
DSS	Diagnostic skills concerning social issues	4 items, each 1–4[b]	Not yet published

[a] 1 = “considerably below average”, 2 = “somewhat below average”, 3 = “average”, 4 = “somewhat above average”, 5 = “considerably above average”

[b] 1 = “strongly disagree”, 2 = “disagree”, 3 = “agree”, 4 = “strongly agree”

2.1.1 Diagnostic Skills (D1–D8)

In COACTIV several established instruments (Hoge & Coladarci, 1989; McElvany et al., 2009; Schrader, 1989) were implemented, targeting different objects of judgment (student achievement vs. motivation; performance on specific tasks vs. the full PISA test) and different levels of judgment (individual students vs. whole class). In the following we will describe the operationalization of eight measures of diagnostic skills[2] ("D1–D8", see Table 1a).

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" **(D1)**; "Please rate the *distribution of achievement* in mathematics in your PISA class relative to an average class of the same school type" **(D2)**; and "Please rate the *motivation* of your PISA class in mathematics relative to an average class of the same school type" **(D5)**. All responses were given on a five-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, teacher responses then were compared with the actual outcomes of their PISA classes. As it is common in the literature on diagnostic competence, small judgment errors (i.e., absolute values of differences) were considered indicators of high diagnostic skills. To this end, we first calculated quantiles for achievement level, distribution of achievement, and motivation 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 objective quintiles, terming the absolute value of this difference the *judgment error* (see Table 1b). Thus, the maximal error was four and a judgment error of zero indicated that the teacher rating was congruent with the objective outcome (the detailed statistical procedure is explicated in Brunner et al., 2013).

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 *low-achievement level* (in the bottom third)" **(D3)** 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 *high-achievement level* (in the top third)" **(D4)**. 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 (see Table 1b).

To evaluate the accuracy of teachers' assessment of task demands, we asked them to estimate how many students in their class would be able to solve each of four specific PISA tasks correctly (A, B, Ca and Cb, see Fig. 2) that addressed important domains of mathematical content typically covered at secondary level

[2] Parts of Sect. 2.1.1 are taken from Brunner et al. (2013).

Table 1b Descriptive results on the aspects of diagnostic skills. For D1–D6: Error = 0 signifies "maximal" diagnostic skill. For D7, D8, QJS, PCK and DSS positive values indicate high performance

Construct E-D1 – E-D6 judgment error of …		Scale	M (SD)
E-D1	Achievement level in PISA test (class average) compared to school type specific German average	0–4[a]	1.20 (0.86)
E-D2	Distribution of achievement (in class)	0–4[a]	1.14 (0.91)
E-D3	% of own students in bottom third of German achievement distribution (in class)	0–100%	0.14 (0.12)
E-D4	% of own students in top third of German achievement distribution (in class)	0–100%	0.22 (0.15)
E-D5	Motivational level (class average) compared to school type specific German average	0–4[a]	1.28 (0.95)
E-D6	% correct solutions with respect to four specific PISA tasks (in class) (task related jugement error)	0–100% (mean error across four tasks)	0.28 (0.11)
D7	Solutions of two specific tasks (kite and Mrs. may) with respect to seven specific students	0–100% (proportion of correct predictions of 2 × 7 = 14 predictions)	0.50 (0.15)
D8	Rank order of achievement of seven specific students in PISA (diagnostic sensitivity)	−1 to 1[b]	0.38 (0.35)
QJS	Quickly classifying student answers as correct or false (12 tasks provided with respective student responses)	0.20–2.38	1.02 (0.36) 12 items $\alpha = 0.71$
PCK	Pedagogical content knowledge	0–35	20.38 (5.71) 22 items $\alpha = 0.78$
DSS	Diagnostic skills concerning social issues	0–4	2.95 (0.47) 4 items $\alpha = 0.88$

Note: M mean, *SD* standard deviation

[a] Estimation within the correct quintile: judgment error of zero. Estimating, for example, the highest quintile, although the lowest quintile would be correct (or vice versa): error of four

[b] Correlation coefficient for ranking (Spearman) between estimated rank order and actual rank order of the seven students

(D6). For each task, we computed the (absolute value of) the 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 (Table 1b). A task-related judgment error of zero again indicates that a teacher correctly estimated the number of correct solutions in his/her PISA class on all four tasks.

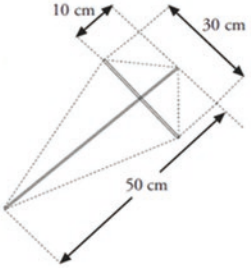
All of the above indicators relate to the class as a whole. To examine the teachers' ability to predict the performance of individual students, we additionally asked

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.

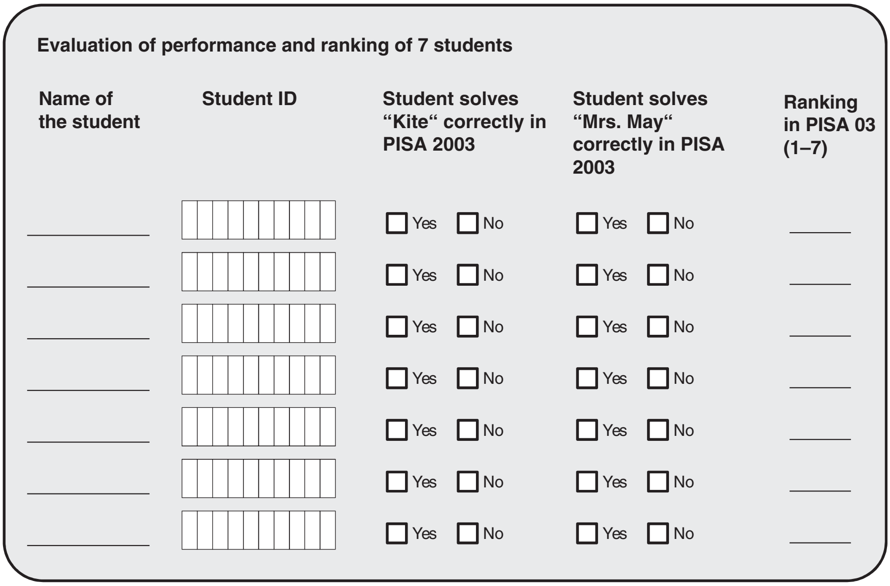
a) Put a cross in the column containing the mistake.

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.

Fig. 2 Specific PISA tasks used in COACTIV to assess teachers’ diagnostic skills (i.e., D6 and D7). All four tasks were provided to the teachers

the teachers to consider seven specific 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 (see Fig. 3). The *accuracy of these individual teacher judgments* **(D7)** was determined by calculating the proportion of the 14 predictions (m = 2 tasks and n = 7 students) 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.



Evaluation of performance and ranking of 7 students

Name of the student	Student ID	Student solves “Kite“ correctly in PISA 2003	Student solves “Mrs. May“ correctly in PISA 2003	Ranking in PISA 03 (1–7)
____________	□□□□□□□□□□	□ Yes □ No	□ Yes □ No	______
____________	□□□□□□□□□□	□ Yes □ No	□ Yes □ No	______
____________	□□□□□□□□□□	□ Yes □ No	□ Yes □ No	______
____________	□□□□□□□□□□	□ Yes □ No	□ Yes □ No	______
____________	□□□□□□□□□□	□ Yes □ No	□ Yes □ No	______
____________	□□□□□□□□□□	□ Yes □ No	□ Yes □ No	______
____________	□□□□□□□□□□	□ Yes □ No	□ Yes □ No	______

Fig. 3 Estimating the performance of seven students in two specific PISA tasks (D7) and estimated ranking of performance of these seven students in the whole PISA test (D8)

Finally, we asked the teachers to judge how well the same seven students probably performed on the whole PISA 2003 mathematics assessment by putting them in rank order of achievement (**D8**, see also Fig. 3). This estimated rank order again was compared with the students’ actual PISA rank order. To provide a measure of *diagnostic sensitivity*, we then computed the rank correlation (Spearman’s Rho) 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.

Thus, the scales of D1–D6 (Table 1a) first had to be transformed into errors E-D1 to E-D6 (Table 1b) and therefore here zero denotes maximal performance, whereas D7 and D8 refer to accuracy or sensitivity itself and therefore here positive values indicate better skills.

2.1.2 Competence Aspects Related to Diagnostic Skills

The following further constructs assessed in COACTIV (an overview on all constructs is provided in the COACTIV-scale documentation, Baumert et al., 2009) are theoretically close to diagnostic skills.

Quick Judgment Skill (QJS)

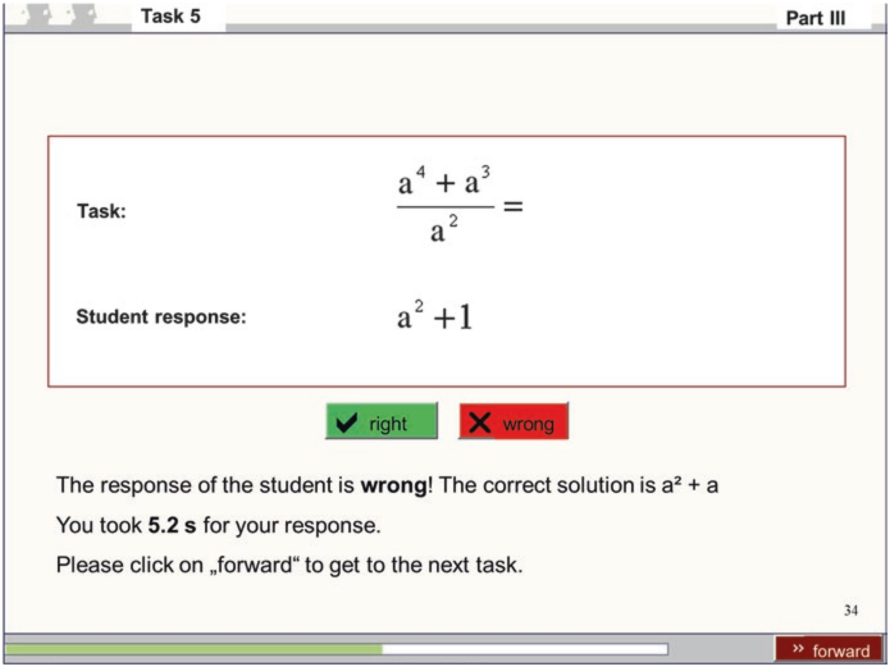
A mathematics teacher should be able to establish the truth or falsehood of students’ statements (or responses to tasks) in mathematics lessons within a reasonable time. Such judgments challenge teachers’ content-specific expertise, because they happen in the publicity of the classroom with all its spontaneity. As mathematics experts teachers should notice failures and they should not take too much time for their identification.

To model this time pressure, in a computer-based instrument 12 easy mathematical tasks were implemented, each with a (hypothetical) corresponding student answer (for a screenshot see Fig. 4). The instruction for the COACTIV teachers was for each task to judge as quickly as possible whether the provided student response was correct or false (all tasks and respective student answers should – without time pressure – constitute no problems for mathematics teachers; all tasks are listed in Krauss & Brunner, 2011). When task and respective student answer were presented simultaneously, the time began to run.

For each teacher and each task the correctness and the time needed for the judgment were recorded. The score for the QJS of a teacher then was calculated by dividing the number of correct judgments (out of 12) by the mean reaction time.

Pedagogical Content Knowledge (PCK)

In COACTIV, two paper and pencil tests on the pedagogical content knowledge (PCK) and the content knowledge (CK) of mathematics teachers were administered (Krauss et al., 2008). Especially the PCK test is of relevance with respect to the present chapter, since two out of three knowledge facets addressed in this test relate



Task: $\frac{a^4 + a^3}{a^2} =$

Student response: $a^2 + 1$

The response of the student is **wrong**! The correct solution is $a^2 + a$

You took **5.2 s** for your response.

Please click on „forward" to get to the next task.

Fig. 4 Sample item of the reaction time test used in COACTIV to assess teachers' QJS (the three lines below appeared after the teacher made his/her decision)

to diagnostic skills. Altogether, PCK was operationalized by 22 items on (for details on the tests and respective results see Krauss et al., 2013):

- Explaining and representing mathematical contents (11 items)
- Mathematics-related student cognitions (typical error and difficulties, 7 items)
- The potential of mathematical tasks (for multiple solution paths, 4 items)

The latter two aspects are closely related to diagnostic skills of mathematics teachers, since they refer to students' thinking and to task properties. Yet – in contrast to D1–D8 – in the PCK test teachers had to answer general questions on typical students' mathematical difficulties and on task properties (such as: "Which problems students *typically* face when …"), i.e., there was no need to relate the answers to actual PISA data.

Diagnostic skills Concerning Social Issues
Diagnostic skills with respect to social issues are another interesting competence facet that might be related to the content-specific skills described so far. In COACTIV, the teacher-scale "diagnostic skills concerning social issues" consists of four items. One sample item was "I notice very quickly, if someone is really sad".

2.1.3 General Model of Diagnostic Competence

Looking at diagnostic skills through various different glasses as in COACTIV is in line with the call of Südkamp and Praetorius (2017), to assess diagnostic competence not in a narrow and constrained sense, but with multiple measures. According to the model of Leuders et al. (2018), which is close to the model of Südkamp and Praetorius (2017), PCK, QJS and DSS are considered diagnostic dispositions, because they were assessed by paper and pencil questionnaires or tests in a laboratory setting outside of the classroom. In contrast, D1–D8 clearly require teachers' perception and interpretation of aspects of their real classes and then to come up with a decision. Yet, because actual decisions were not observed in the real classroom context (but again in the laboratory setting), we theoretically subsume D1–D8 under diagnostic thinking (middle column of Fig. 5).

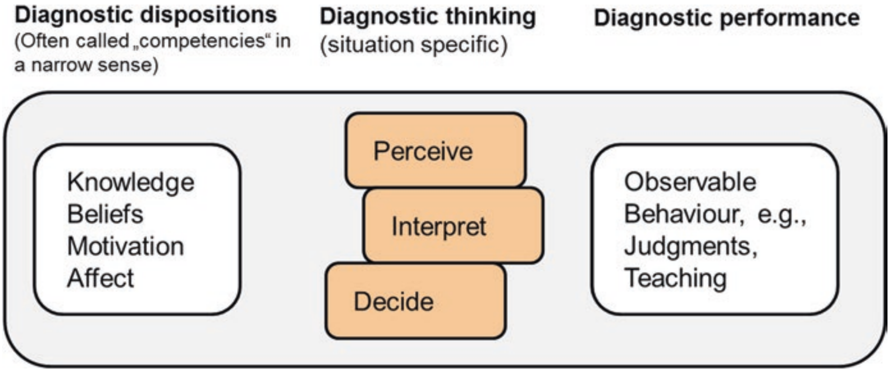


Fig. 5 Diagnostic competence (in a wider sense) as a continuum (Leuders et al., 2018)

2.2 Instructional Quality and Student Characteristics

In order to assess teachers' instructional quality in COACTIV, a parsimonious model with three latent dimensions, which are each represented by multiple indicators, was developed (Fig. 6, for details see Kunter and Voss, 2013). Very briefly, the potential for *cognitive activation* was assessed in terms of the cognitive quality of the mathematical tasks implemented by the teachers in class tests (e.g., the need for mathematical argumentation; cf. Kunter et al., 2013). Class tests were chosen because they allow valid conclusions to be drawn about the intended purposes of instruction. The dimension of *classroom management* was assessed using scales from both the student (PISA) and the teacher (COACTIV) questionnaires asking, for instance, for disruption levels or time wasted. Indicators of *individual learning support* were formed by scales from the student questionnaire, assessing various aspects of the interaction between students and teachers (see Kunter & Voss, 2013). The students' learning gain was estimated by the mathematical achievement in PISA 2004 (while controlling for the achievement in the preceding year, i.e. PISA 2003). The full mediation model is explicated in Figs. 6 and 8.

In Sect. 3.4 structural equation modeling will be conducted in order to estimate the predictive validity of aspects of diagnostic skills for lesson quality and students' learning gains (structural equation models with respect to various other teacher competence aspects as predictors are summarized in Kunter, Baumert et al., 2013, Kunter, Klusmann et al., 2013, or in Krauss et al., 2017).

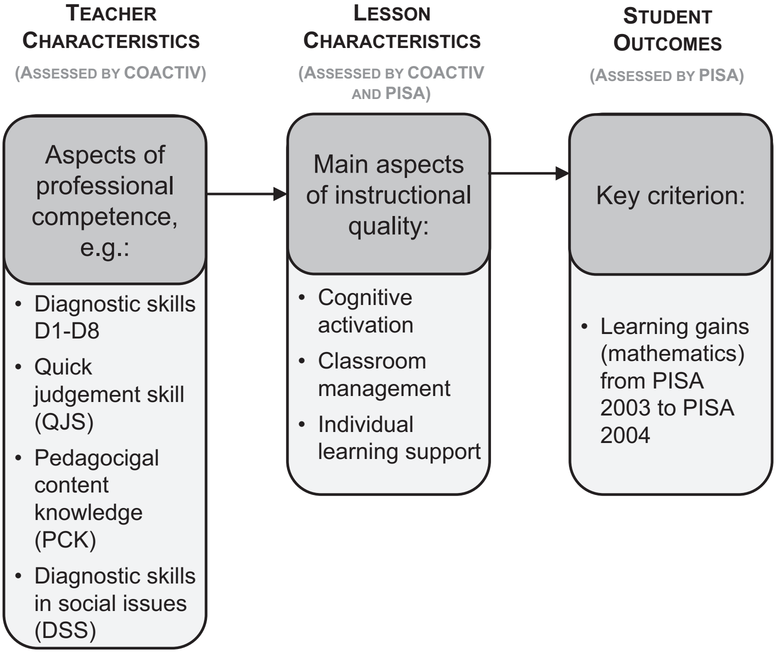

TEACHER CHARACTERISTICS
(ASSESSED BY COACTIV)
LESSON CHARACTERISTICS
(ASSESSED BY COACTIV AND PISA)
STUDENT OUTCOMES
(ASSESSED BY PISA)
Aspects of professional competence, e.g.:
• Diagnostic skills D1-D8
• Quick judgement skill (QJS)
• Pedagocigal content knowledge (PCK)
• Diagnostic skills in social issues (DSS)
Main aspects of instructional quality:
• Cognitive activation
• Classroom management
• Individual learning support
Key criterion:
• Learning gains (mathematics) from PISA 2003 to PISA 2004

Fig. 6 Causal model in COACTIV

So far, empirical evidence on the predictive validity of aspects of diagnostic competence for student learning is mixed (see e.g., Gabriele, Joram, & Park, 2016). Several studies document predictive validity, some studies show moderation by instructional variables and other studies find no predictive evidence. It can be assumed that the precise operationalization and measurement of diagnostic competence are essential and that the respective variation might explain the differing empirical results at least partially (e.g., Artelt & Rausch, 2014).

3 Results

In Sect. 3.1, we report descriptive results with respect to the constructs assessed and in Sect. 3.2 we analyze the relationship between the different diagnostic skills. As previous research has shown major differences between teachers' competences with respect to different German secondary school types (Kunter, Baumert et al., 2013), we also report respective differences in diagnostic skills in Sect. 3.3. Finally, in Sect. 3.4, we conduct structural equation modeling in order to analyze the effects of diagnostic skills on instructional quality and on learning gains of students.

3.1 Descriptive Results

To estimate teachers' diagnostic skills with respect to the constructs displayed in Table 1a, judgment errors regarding the skills D1–D6 were calculated (Table 1b). While D1–D5 and D8 depend on single item measures, QJS, PCK and DSS consist of multi-item-scales (therefore for the latter three constructs, Cronbach's alpha is provided in Table 1b). Since the scales of E-D1, E-D2 and E-D5 share the same range, judging achievement level, achievement distribution and motivational level obviously was similarly difficult (yielding means of 1.18, 1.22 and 1.30, see Table 1b).

With respect to D7 it should be noted that the empirical mean almost perfectly represents the probability of guessing. Obviously it is difficult for teachers to predict the performance of individual students in certain tasks. The mean of QJS was 1.02, because on average teachers judged 8.8 of the 12 items correctly and on average needed 9.7 s for each judgment (including reading the item).

3.2 Relationship Between Diagnostic Skills

Table 2 shows the intercorrelations between the various indicators of diagnostic skills for the whole sample (in each cell above) and for the academic and non-academic track separately (the two values in each cell below)[3]. In Germany there are

[3] The intercorrelations with QJS, PCK and DSS as well as the school type related correlations were not reported in Brunner et al. (2013).

basically three secondary school types, namely *Gymnasium* (academic track, prerequisite for the admission to university), *Realschule* (intermediate track) and *Hauptschule* (vocational track). The respective teacher education differs between Gymnasium (higher proportion of content courses) and the latter two school types (higher proportion of educational courses). Therefore in COACTIV analyses usually the performance of the academic track teachers (GY) is compared with teachers of the other tracks, which are called Non-Gymnasium (NGY). For details concerning the German school system see, for example, Cortina and Thames (2013).

In Table 2, E-D1 to E-D6 denote errors and D7 to DSS denote the competences themselves (displayed within the dottet rectangle in Table 2). Thus, theoretically the correlations of E-D1 to E-D6 and of D7 to DSS should be positive, while each of the errors should correlate negatively with each of the competences.

As it becomes clear from Table 2, there seems to be no systematic pattern within the bivariate correlations. Because there is no appearance of a dominant dimension of diagnostic competence, the constructs D1 to D8 were named “diagnostic skills” by Brunner et al. (2013). This pattern of results – only weak or no correlations between different indicators of diagnostic skills – was also reported by both Schrader (1989) and Spinath (2005).

However, judging the achievement level (D1) and judging the bottom third (D3) and the top third of achievement distribution (D4) seem to be correlated as well as the skills D6, D7 and D8. Interestingly, DSS is even associated negatively with some other competence aspects. It seems to be that teachers, who concentrate on social aspects of students, are less competent with respect to, for example, D3, D8 and PCK. Furthermore, there are highly differential correlations of D6 and QJS: While in the group of NGY-teachers the correlations are significantly negative, in the GY-group the opposite is true.

However, the results in Table 2 should be interpreted with caution, because only about half of the correlations point in the expected direction. Of course the different measurements of diagnostic skills can be criticized. Perhaps constructing quintiles, for instance, may not be the best procedure to judge D1, D2 and D5. Furthermore, estimating students’ performance in the whole PISA test might be difficult, because teachers did not know all items of this test. However, Table 2 corroborates the assumption that measuring judgment accuracies might highly depend on the exact operationalization (Gabriele et al., 2016).

3.3 School-Type Differences with Regard to the Mean Levels

In Table 3, mean levels of the teachers of the academic track (GY) are compared with the non-academic track teachers (NGY). There were differences with respect to school type in favor of GY-teachers, especially regarding D4, D6, QJS and PCK. Interestingly, the diagnostic skills with respect to social issues are descriptively more pronounced in NGY-teachers.

Table 2 Intercorrelations of the indicators of diagnostic skills (Pearson product–moment correlation) for all teachers (upper half in each cell) and for teachers of academic (GY) and non-academic track (NGY) separately (lower half)

All	E-D1		E-D2		E-D3		E-D4		E-D5		E-D6		D7		D8		QJS		PCK		DSS	
GY NYG																						
E-D1	--																					
E-D2	0.12		--						all teachers													
	0.23	0.01							GY	NGY												
E-D3	**0.18**		-0.10		--																	
	0.06	**0.26**	-0.06	-0.12																		
E-D4	***0.29****		0.11		-0.09		--															
	0.21	***0.34***	0.18	0.08	***-0.30***	0.00																
E-D5	-0.05		-0.02		-0.08		-0.09		--													
	-0.09	-0.02	-0.12	0.06	0.03	-0.15	-0.22	-0.04														
E-D6	-0.01		0.03		0.00		0.08		0.07		--											
	0.01	-0.03	-0.01	0.06	0.01	-0.05	-0.13	0.10	0.09	0.05												
D-7	-0.04		-0.05		0.14		-0.09		-0.11		***-0.33***		--									
	-0.05	0.01	0.13	**-0.23**	0.19	0.11	0.03	-0.12	-0.16	-0.07	-0.19	***-0.40***										
D-8	0.00		0.03		-0.07		0.03		0.07		**-0.16**		**0.17**		--							
	0.01	-0.02	0.15	-0.09	0.01	-0.13	0.04	0.02	0.13	0.01	-0.23	-0.14	0.18	0.17								
QJS	0.03		0.06		-0.03		-0.01		-0.09		-0.14		0.09		0.00		--					
	0.12	-0.05	0.08	0.04	0.06	-0.07	0.02	0.05	-0.03	-0.12	0.20	**-0.27**	-0.05	0.19	-0.07	0.09						
PCK	0.06		0.10		-0.14		-0.08		-0.14		-0.08		*-0.07*		0.13		0.13		--			
	0.09	0.05	-0.01	0.19	-0.13	-0.11	0.08	-0.10	-0.20	-0.09	0.05	0.01	-0.00	-0.21	-0.04	**0.27**	-0.03	0.11				
DSS	-0.01		-0.06		**0.17**		-0.01		0.03		0.07		-0.09		**-0.17**		0.02		**-0.20**		--	
	0.13	-0.14	0.00	-0.11	**0.28**	***0.09***	-0.13	0.05	-0.08	0.13	0.14	-0.02	-0.10	-0.06	***-0.30***	-0.05	-0.02	0.09	-0.14	**-0.22**		

$N_{All,\ min}$=136
$N_{All,\ max}$=180

$N_{GY,\ min}$=62
$N_{GY,\ max}$=77

$N_{GY,\ min}$=74
$N_{GY,\ max}$=103

Bold**: p≤0.05, ***Bold and italics: p≤0.01

Table 3 Differences in diagnostic skills by school type

Dimension of (sub-)skills	Academic track GY N M (SD)	Non-academic track NGY N M (SD)	Group differences	
			d	*p*-value
E-D1 achievement level in PISA-test (class average) compared to school type specific German average	*N* = 103 1.12 (0.94)	*N* = 78 1.20 (0.86)	0.09	0.51
E-D2 distribution of achievement (in class)	*N* = 103 1.13 (1.03)	*N* = 77 1.14 (0.91)	0.01	0.97
E-D3% of own students in bottom third of German achievement distribution (in class)	*N* = 102 0.13 (0.11)	*N* = 75 0.14 (0.12)	0.09	0.49
E-D4% of own students in top third of German achievement distribution (in class)	*N* = 101 0.17 (0.11)	*N* = 75 0.22 (0.15)	0.38	0.02
E-D5 motivational level (class average) compared to school type specific German average	*N* = 102 1.23 (0.87)	*N* = 77 1.28 (0.95)	0.05	0.72
E-D6% correct solutions with respect to four specific PISA tasks in class (task related judgment error)	*N* = 88 0.22 (0.08)	*N* = 66 0.28 (0.11)	0.63	<0.01
D7 solutions of two specific tasks (Kite and Mrs. May) with respect to seven specific students	*N* = 91 0.54 (0.15)	*N* = 74 0.50 (0.15)	0.27	0.10
D8 rank order of achievement of seven specific students in PISA (diagnostic sensitivity)	*N* = 91 0.37 (0.37)	*N* = 74 0.38 (0.35)	−0.03	0.93
QJS quickly classifying student answers as correct or false (12 tasks provided with respective student responses)	*N* = 87 1.13 (0.38)	*N* = 71 0.94 (0.33)	0.54	<0.01
PCK pedagogical content knowledge	*N* = 94 22.50 (5.43)	*N* = 73 18.46 (5.68)	0.72	<0.01
DSS diagnostic skills concerning social issues	*N* = 99 2.88 (0.53)	*N* = 77 2.95 (0.47)	−0.14	0.36

Note*:* *M* mean, *SD* standard deviation, *d* effect size according to Cohen (1992). The effect sizes in Table 3 were always calculated in a way, so that *positive effect sizes mean advantage of the GY-teachers* (already acknowledging that lower errors denote higher performances)

These results mirror previous COACTIV results since we also found effects in favor of GY-teachers with respect to many other *content-related* competence aspects and effects in favor of NGY-teachers regarding some further *non-content-related* competences in COACTIV (Kunter, Baumert et al., 2013).

3.4 Predictive Validity with Respect to Teaching Quality and Student Learning

Despite all problems concerning reliability and validity of the constructs D1 to D8 reported above, we ran a series of two-level structural equation models to tentatively check the predictive validity of the constructs assessed. Since the large predictive validity of PCK was previously documented in various black box and in mediation models (Baumert & Kunter, 2013; Baumert et al., 2010; Kunter, Klusmann et al., 2013), this construct will be ignored in the following (for further information on objectivity, reliability and validity of QJS or PCK, see Krauss & Brunner, 2011 or Krauss et al., 2013, respectively).

First, we specified nine separated black box models (see Fig. 7 or Table 4) where D1–D8 and QJS should predict student achievement in PISA 2004. At the class level, we controlled for school type (GY vs. NGY) and on the individual level we controlled for prior knowledge in mathematics (PISA achievement in 2003), reading literacy, basic cognitive abilities, immigration status and socio-economic status

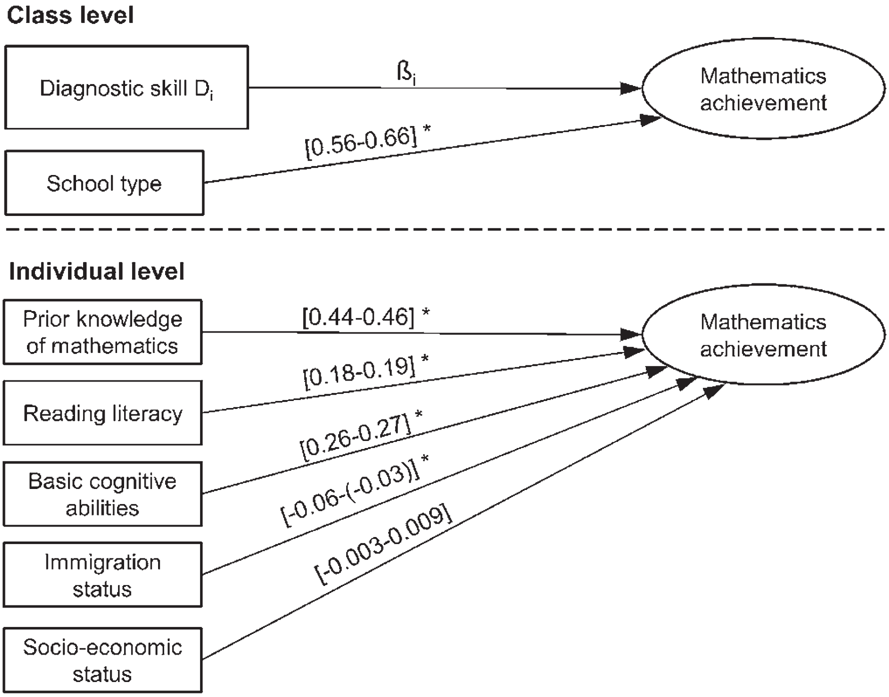


Fig. 7 Black box models with the predictors D1–D8 and QJS and the criterion students' mathematics achievement (the single results for the standardized regression coefficients β_i are depicted in Table 4); *$p<0.05$

Table 4 Nine black box models and the respective standardized regression coefficients β for Fig. 7 (criterion: mathematics achievement)

Model	Predictor	β	p-value
Model 1 E-D1	E-D1 achievement level in PISA test (class average)	0.28*	<0.01
Model 2 E-D2	E-D2 distribution of achievement (in class)	0.03	0.81
Model 3 E-D3	E-D3% students in bottom third of achievement distribution (in class)	−0.14	0.16
Model 4 E-D4	E-D4% students in top third of achievement distribution (in class)	0.27*	<0.01
Model 5 E-D5	E-D5 motivational level (class average)	−0.03	0.77
Model 6 E-D6	E-D6% correct solutions with respect to 4 specific PISA tasks (in class)	−0.21	0.06
Model 7 D7	D7 solutions of 2 specific tasks (kite and Mrs. may) with respect to seven specific students	0.07	0.53
Model 8 D8	D8 rank order of these seven specific students in PISA	−0.02	0.83
Model 9 QJS	QJS quickly classifying student answers as correct or false (12 tasks with respective student responses)	−0.03	0.80

β Standardized regression coefficient, *$p<0.05$

(in brackets in Fig. 7 the range of the corresponding standardized regression coefficients with respect to all nine models is depicted).

Table 4 summarizes the results of all nine black box models. It should be noted that the significant positive coefficients in model 1 and model 4 denote *negative effects*, because with respect to D1 and D4 errors were modeled (for an explanatory attempt, see the respective mediation models later). The only model that is close to a *positive effect* on student achievement is model 6 (judging the proportions of correct solutions in four specific PISA tasks in the class). Anders et al. (2010) demonstrated that this effect becomes significant if only the two tasks “Sausage Stand a” and “Sausage Stand b” are analyzed (maybe because proportionality is an intensively treated topic in mathematics in Germany). We could not replicate the positive effect of D8 from Anders et al. (2010) because of a different composition of the teacher sample analyzed.

In a second series we implemented nine corresponding mediation models[4] (Fig. 8). The measurement of the instructional quality by the latent constructs cognitive activation (assessed by cognitive level of tasks), learning support and classroom management was previously described in Sect. 2.2. We found that the negative effects of D1 and D4 were mediated in both cases by a positively significant

[4] In the tradition of COACTIV we name these models “mediation models” (Baumert et al., 2010, Kunter, Klusmann et al., 2013), although the use of this term usually implies addressing, e.g., the multiplicative term of the indirect path, etc.

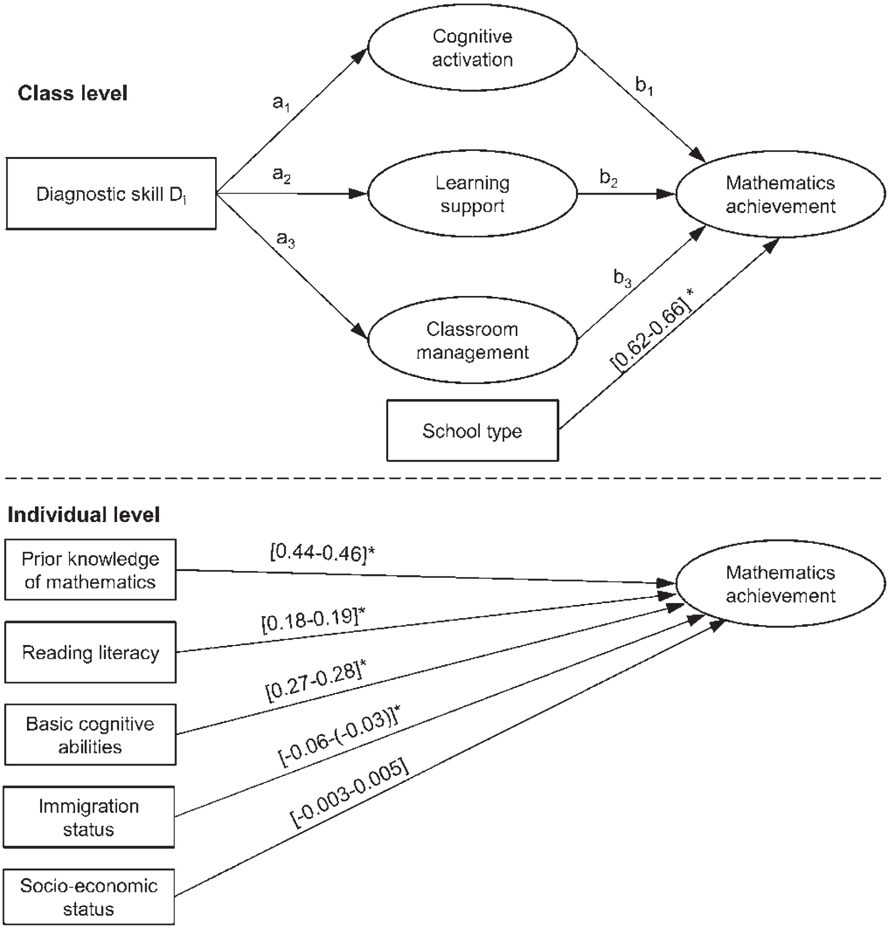


Fig. 8 Overview of the nine two-level "mediation" models (*Note*: a_i and b_i denote the respective standardized regression coefficients, D1-D8: predictors, *p<0.05)

path a_1 (see Fig. 8), which means that teachers with *less* D1 and D4 implemented *more* cognitively demanding tasks in their classes. This is an interesting finding since – *although,* or perhaps even *because* of misjudging (!) the achievement level – teachers dare to implement cognitively activating tasks, which in turn leads to higher mathematics achievement.

The only significant effect we found with respect to a_2 and a_3 was in model 8, were a_2 was positively significant and a_3 was negatively significant, indicating that D8 (estimating the rank order of seven students) works differently than the other predictors. Concerning b_1 to b_3 in almost all models b_1 and b_3, were significant while b_2 was not (which replicated the results of other structural equation models of COACTIV, see, e.g., Kunter, Klusmann et al., 2013).

4 Discussion

In the present chapter, we summarized the findings of Anders et al. (2010) and Brunner et al. (2013) and added the constructs of quickly judging student responses (QJS), pedagogical content knowledge (PCK) and diagnostic skills concerning social issues (DSS), which – at least theoretically – should be close to diagnostic skills. However, correlational analyses yielded only moderate and even partially unexpected results. In contrast, mean level differences between Gymnasium- and Non-Gymnasium-teachers are in line with previous COACTIV results and thus (cautiously) validate the constructs implemented.

Finally, we implemented structural equation models to assess the impact of diagnostic skills and QJS directly on students' mathematical achievement (black box models) and models where this assumed effect was mediated by central aspects of instructional quality. Interestingly, the precise judgment of student achievement level may even prevent teachers from implementing cognitively demanding mathematical tasks (mediation models 1 and 4). Only the correct judgment of the solution rate with respect to four specific tasks (that explicitly were shown to the teachers) seems to have a positive impact on students' achievement (black box model 6). Note that estimating the performance of their class in the whole PISA test obviously was difficult for teachers, maybe because they do not know the concrete tasks implemented in PISA.

Taken together, our analyses confirm the use of the term "diagnostic skills" (instead of diagnostic competence) by Brunner et al. (2013), because we found only unsystematic and moderate relationships between the constructs analyzed. Our results are in line with Spinath (2005), who also found only weak or no correlations between different indicators of diagnostic skills. The present chapter, however, is far from stating final conclusions, but aims to introduce and describe various ways to theoretically and empirically examine diagnostic skills of teachers in the subject of mathematics.

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. *Psychologie in Erziehung und Unterricht*, *57*, 175–193.

Artelt, C., & Rausch, T. (2014). Accuracy of teacher judgments: When and for what reasons? In S. Krolak-Schwerdt, S. Glock, & M. Böhmer (Eds.), *Teachers' professional development: Assessment, training, and learning* (pp. 229–248). Rotterdam, The Netherlands: Sense Publishers.

Baumert, J., Blum, W., Brunner, M., Dubberke, T., Jordan, A., Klusmann, U., Krauss, S., Kunter, M., Löwen, K., Neubrand, M., & Tsai, Y.-M. (2009). *Professionswissen von Lehrkräften, kognitiv aktivierender Mathematikunterricht und die Entwicklung von mathematischer Kompetenz (COACTIV): Dokumentation der Erhebungsinstrumente* (Materialien aus der Bildungsforschung, 83). Berlin, Germany: Max-Planck-Institut für Bildungsforschung.

Baumert, J., Kunter, M., Blum, W., Brunner, M., Voss, T., Jordan, A., Klusmann, U., Krauss, S., Neubrand, M., & Tsai, Y.-M. (2010). Teachers' mathematical knowledge, cognitive activation in the classroom, and student progress. *American Educational Research Journal, 47*(1), 133–180.

Baumert, J., & Kunter, M. (2013). The COACTIV model of teachers' professional competence. In M. Kunter, J. Baumert, W. Blum, U. Klusmann, S. Krauss, & M. Neubrand (Eds.), *Cognitive activation in the mathematics classroom and professional competence of teachers, mathematics teacher education* (Vol. 8, pp. 25–48). New York, NY: Springer.

Brunner, M., Anders, Y., Hachfeld, A., & Krauss, S. (2013). The diagnostic skills of mathematics teachers. In M. Kunter, J. Baumert, W. Blum, U. Klusmann, S. Krauss, & M. Neubrand (Eds.), *Cognitive activation in the mathematics classroom and professional competence of teachers mathematics teacher education* (Vol. 8, pp. 229–248). New York, NY: Springer.

Cohen, J. (1992). A power primer. *Psychological Bulletin*, *112*(1), 155.

Cortina, K. S., & Thames, M. H. (2013). Teacher education in Germany. In M. Kunter, J. Baumert, W. Blum, U. Klusmann, S. Krauss, & M. Neubrand (Eds.), *Cognitive activation in the mathematics classroom and professional competence of teachers, mathematics teacher education* (Vol. 8, pp. 49–62). New York, NY: Springer.

Gabriele, A. J., Joram, E., & Park, K. H. (2016). Elementary mathematics teachers' judgment accuracy and calibration accuracy: Do they predict students' mathematics achievement outcomes? *Learning and Instruction*, *45*, 49–60.

Hoge, R. D., & Coladarci, T. (1989). Teacher-based judgments of academic achievement: A review of literature. *Review of Educational Research*, *59*(3), 297–313.

Krauss, S., Blum, W., Brunner, M., Neubrand, M., Baumert, J., Kunter, M., et al. (2013). Mathematics teachers' domain-specific professional knowledge: Conceptualization and test construction in COACTIV. In M. Kunter, J. Baumert, W. Blum, U. Klusmann, S. Krauss, & M. Neubrand (Eds.), *Cognitive activation in the mathematics classroom and professional competence of teachers, mathematics teacher education* (Vol. 8, pp. 147–174). New York, NY: Springer.

Krauss, S., Brunner, M., Kunter, M., Baumert, J., Blum, W., Neubrand, M., & Jordan, A. (2008). Pedagogical content knowledge and content knowledge of secondary mathematics teachers. *Journal of Educational Psychology, 100*(3), 716–725.

Krauss, S., Lindl, A., Schilcher, A., Fricke, M., Göhring, A., Hofmann, B., Kirchhoff, P. & Mulder, R. H. (Hrsg.). (2017). *FALKO: Fachspezifische Lehrerkompetenzen. Konzeption von Professionswissenstests in den Fächern Deutsch, Englisch, Latein, Physik, Musik, Evangelische Religion und Pädagogik*. Münster: Waxmann.

Krauss, S., & Brunner, M. (2011). Schnelles Beurteilen von Schülerantworten: Ein Reaktionszeittest für Mathematiklehrer/innen. *Journal für Mathematik-Didaktik*, *32*(2), 233.

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

Kunter, M., Klusmann, U., Baumert, J., Richter, D., Voss, T., & Hachfeld, A. (2013). Professional competence of teachers: Effects on instructional quality and student development. *Journal of Educational Psychology*, *105*(3), 805–820. https://doi.org/10.1037/a0032583

Kunter, M., & Voss, T. (2013). The model of instructional quality in COACTIV: A multicriteria analysis. In M. Kunter, J. Baumert, W. Blum, U. Klusmann, S. Krauss, & M. Neubrand (Eds.), *Cognitive activation in the mathematics classroom and professional competence of teachers, mathematics teacher education* (Vol. 8, pp. 97–124). New York, NY: Springer.

Leuders, T., Dörfler, T., Leuders, J., & Philipp, K. (2018). Diagnostic competences of mathematics teachers – Unpacking a complex construct. In K. Philipp, T. Leuders, & J. Leuders (Eds.), *Diagnostic competences of mathematics teachers*. New York, NY: Springer.

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. *Zeitschrift für pädagogische Psychologie*, *23*(34), 223–235.

Prenzel, M., Baumert, J., Blum, W., Lehmann, R., Leutner, D., Neubrand, M., et al. (Eds.). (2004). *PISA 2003. Der Bildungsstand der Jugendlichen in Deutschland—Ergebnisse des zweiten internationalen Vergleichs*. Münster, Germany: Waxmann.

Schrader, F. W. (1989). *Diagnostische Kompetenzen von Lehrern und ihre Bedeutung für die Gestaltung und Effektivität des Unterrichts*. Frankfurt am Main, Germany: Lang.

Spinath, B. (2005). Akkuratheit der Einschätzung von Schülermerkmalen durch Lehrer und das Konstrukt der diagnostischen Kompetenz: Accuracy of teacher judgments on student characteristics and the construct of diagnostic competence. *Zeitschrift für pädagogische Psychologie*, *19*(1/2), 85–95.

Südkamp, A., & Praetorius, A. K. (Eds.). (2017). *Diagnostische Kompetenz von Lehrkräften: Theoretische und methodische Weiterentwicklungen*. Waxmann Verlag.