

Perception of student errors under time limitation: Are teachers faster than mathematicians or students?

Lena Pankow (University of Hamburg), Gabriele Kaiser (University of Hamburg, Australian Catholic University), Johannes König (University of Cologne), Sigrid Blömeke (University of Oslo, CEMO)

Keywords: validation study; contrast groups; timed test; knowledge facets; student error recognition

Abstract

The ability to offer constructive feedback to student errors is an indispensable requirement for mathematics teachers for providing cognitively challenging learning opportunities. However, if they are to react adequately, teachers need to identify student errors immediately. The fast perception of student errors can therefore be described as an indispensable part of mathematics teachers' professional competence. Data on this facet of teacher competence was gathered as part of a national follow-up-study of the IEA's international TEDS-M (Teacher Education and Development Study in Mathematics) that used a time-limited test to measure teachers' perception of student errors. This paper aims to provide evidence for the validity of the test interpretation of fast student-error perception as an indicator of professional competence by using contrast groups already used in other studies. Overall, the study could support the validity of the test interpretation because the chosen contrast groups are found to perform either better than the tested teachers – as is the case for the contrast group of mathematicians – or poorer, as is the case for the group of students. Furthermore, the present study shows that the competence facet of fast student error perception is closer to the domain of teachers' mathematical content knowledge than it is to the domain of teachers' mathematics pedagogical content knowledge.

1. Introduction

Perceiving and adequately responding to student errors is part of mathematics teachers' daily activities in the classroom. Teachers do not only have to be aware of student errors; they must also identify, interpret and classify them. Further, they must decide if the error is relevant enough to be addressed extensively in front of the whole group or individually, and whether the error is caused by an underlying misconception. The time they have for noticing and addressing student errors is limited, but doing so, according to a study by Schoy-Lutz (2005), creates genuine learning opportunities in most cases. A necessary condition for recognizing these learning opportunities is the fast identification of student errors.

Knowledge of students' misconceptions is an important facet of teacher competence in mathematics (cf. Altmann and Nückles 2017). This competence was assessed in TEDS-FU, a German follow-up study to the international Teacher Education and Development Study in Mathematics (TEDS-M; for details see Blömeke et al. 2014). TEDS-FU evaluated the competence of a subsample of German TEDS-M participants who were practicing mathematics teachers in their fourth year of teaching. The results presented in this paper are related to lower secondary teachers. TEDS-FU contained, in addition to the original TEDS-M

study, a test component on fast perception of student errors by teachers (see Pankow et al. 2016). The fast perception of students' mathematical errors can be described as a sub-facet of *diagnostic competence* (Shulman, 1986).

To validate the interpretation of the test as an indicator of mathematics teachers' competence and to further examine whether the instrument is testing the targeted construct, the present study includes a contrasting group analysis based on five groups with varying expertise: future mathematics teachers in their master study (Sample 1), early career teachers (about three to four years of teaching experience) (Sample 2), experienced teachers (>4 years of teaching experience) (Sample 3), university students enrolled in a mathematics degree program (Sample 4), and high school students (Sample 5). This approach was adopted from the COACTIV (Professional Competence of Teachers, Cognitively Activating Instruction, and Development of Students' Mathematical Literacy) study (cf. Krauss, Baumert and Blum 2008), in which the tests on mathematical content knowledge and mathematics pedagogical content knowledge – including a test of fast perception of student errors – were validated using contrast groups. Like the ones in the COACTIV study, the groups in this study were chosen according to their different levels and qualities of expertise, i.e., mathematical content knowledge and mathematics pedagogical content knowledge. Therefore, this study aims to examine the relations between a key dependent variable – fast perception of student error – and a set of external variables, an approach that can be considered as part of criterion-based *validation*, at least when accounting for a broader understanding of the classical criterion validity concept. Furthermore, the relation between the content of the test and the construct it is intended to measure is analyzed, providing content-oriented evidence that, in a classical sense, could be labelled as *construct validity* (AERA, APA, and NCME 2014).

After a description of the theoretical background of the study, which includes an overview of the current discussion on student errors in mathematics education, the construct of fast perception of student errors under time pressure is developed. Furthermore, the validation process using the contrast group method is described along with other sources of evidence for validity. Finally, the roles and measurement of both mathematical content knowledge and mathematics pedagogical content knowledge within the TEDS-FU study are discussed. After developing the research aim and the hypotheses, the design of the validation study is presented, followed by a description of the results. The paper closes with a summary and discussion of those results.

2. Theoretical framework of the study

This section describes the theoretical framework of student errors and teachers' familiarity with those errors as part of teachers' diagnostic competence. That discussion is followed by a description of the empirical studies that operationalize and measure student-error perception. The chapter closes with a description of different approaches to the validation of the test instrument.

2.1. Dealing with student errors

Oser et al. (1999) defined an error as a “deviation from the standard. Standards represent the reference system; without standards or rules, it would not be possible to distinguish between correct and incorrect” (p. 11). Heinze (2004) specified this general definition of an error

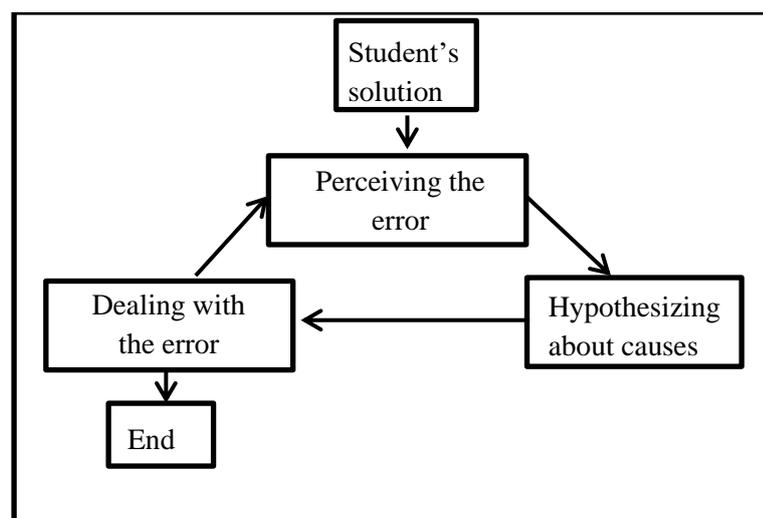
tailored to mathematics education: “An error is an expression which is against the general statements and definitions of mathematics as well as against a generally accepted mathematical-methodical approach” (p. 223). Student errors in mathematics can happen for a variety of reasons, including carelessness, ignoring the given rules, or feeling uncertain. However, many student errors are of a systematic nature – i.e., they will occur again in another mathematical task of equivalent structure. If these types of errors are not tackled adequately by a teacher’s intervention (Swan 2004), they will probably be repeated over subsequent school years (Radatz 1980, p. 16; Türling 2014). The ability to quickly perceive student errors represents an important part of professional competence, then, because immediate feedback from a teacher supports student learning processes and enhances active learning. In addition, the perception of student errors is needed in order to evaluate student work and plan the next instructional steps (Brühwiler 2017).

2.2. Student-error perception

The conceptualization of teacher competence and its measurement has been the focus of much empirical research in the last two decades. However, the inclusion of student-error perception as part of that measurement, and the ways of dealing with that component, have been introduced into that discussion only recently, especially in the context of adaptive teacher behavior (Südkamp and Praetorius 2017).

Departing from the general discussion of diagnostic competencies as part of teacher competence and how teacher competence can be fostered holistically as part of teacher education, Heinrichs and Kaiser (2018) developed a model for the perception and handling of errors in mathematics instruction. They described the diagnostic process in instructional situations, where errors occur by way of a cyclic process starting from the perception of a student error, leading to the development of hypotheses of the causes of these errors, and ending with possible instructional approaches for dealing with the error in class (see also Hoth 2016).

Fig. 1: Model of the diagnostic process in error situations (Heinrichs and Kaiser 2018, p. 84)



If the student has not understood the error and therefore overcome the problem, the cyclic process has to be carried out another time (see Fig. 1). The first step of the diagnostic process, the perception of the error, was the focus of the time-limited test presented in this paper.

Neither of the other steps of the diagnostic process – ‘hypothesizing about causes’ and ‘dealing with the error’ – were part of the test developed in TEDS-FU.

Based on the current discussion of teacher competence, Blömeke, Gustafsson and Shavelson (2015) described competence as a continuum, departing from teachers’ knowledge as their dispositions, including situational components such as perception, interpretation and decision-making, leading to the observable behavior of teachers in class (Blömeke et al., 2015, p. 7). As error perception may be considered to be embedded within the perception component seen here as part of the situational competences of teachers, it can, in a narrower sense, also be seen as a component of diagnostic competence. Diagnostic competence is connected to mathematical content knowledge and mathematics pedagogical content knowledge as this kind of knowledge is needed for the perception and interpretation of student errors as well as for the development of adequate teaching measures (for an overview of the current discussion on diagnostic competence see Südkamp and Praetorius 2017, Leuders et al. 2018). A crucial element of student-error perception is the speed of error perception under time constraints, as teachers have to analyze students’ statements and responses within a limited time frame (Lindmeier, Heinze and Reiss 2013, p. 106); dealing with those time constraints is part of teachers’ professional lives (Wahl, Weinert and Huber 1984).

The knowledge dimensions of teachers’ professional competence are conceptualized in TEDS-FU based on the work of Shulman (1986), who distinguished amongst others content knowledge, pedagogical content knowledge and pedagogical respectively curricular knowledge. In the context of the test of fast perception of student errors, two of Shulman’s three facets are relevant, namely *content knowledge* as the part of knowledge that encloses the content and is not reduced, or pure factual knowledge. As this facet relates to mathematics, it is called *Mathematical Content Knowledge* (MCK). Furthermore, Shulman (1986) hones in on specific pedagogical knowledge, the *Pedagogical Content Knowledge* (PCK), which includes the following:

. . . for the most regularly taught topics in one’s subject area, the most useful forms of representation of those ideas, the most powerful analogies, illustrations, examples, explanations, and demonstrations – in a word, the ways of representing and formulating the subject that make it comprehensible to others. (Shulman 1986, p. 9)

This form of knowledge translates to mathematics as *Mathematical Pedagogical Content Knowledge* (MPCK). Additionally, Shulman points out that the teacher who is confronted with student errors has to have strategies to correct misconceptions in order to reorganize students’ knowledge (Shulman 1986, pp. 9).

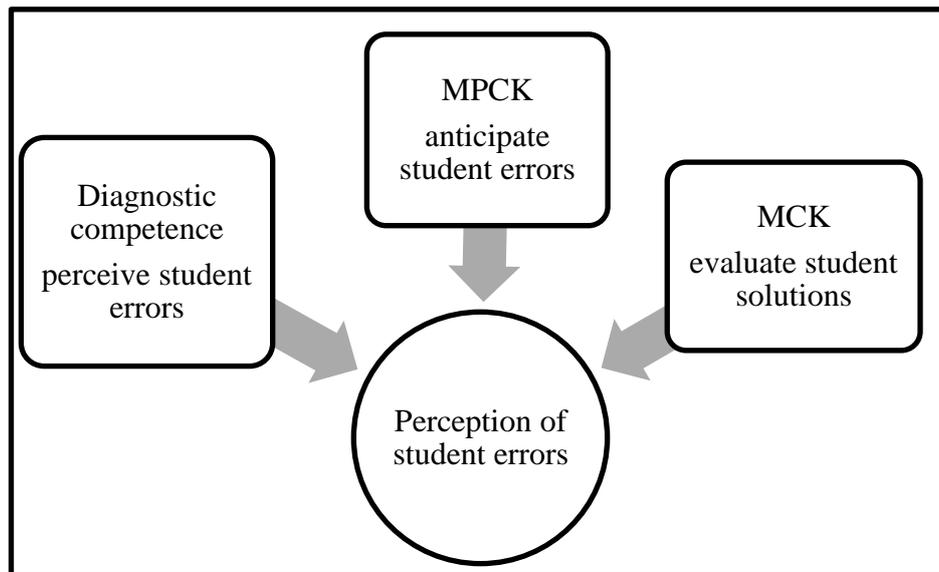


Fig. 2: Contribution of different theoretical constructs to the perception of students' errors

Overall, the perception of student errors has different theoretical starting points as described in Figure 2. The initial step in diagnosing a student error is the perception of a deviation from the expected norm. Reisman (1976) describes this phase as *identification* as it refers to noticing and analyzing student behavior. To use errors and the misconceptions connected to them productively in the learning process, it is indispensable to identify them quickly (cf. Leuders 2001; Radatz 1980a, 1980b).

2.3. Empirical studies on fast student-error perception

Teacher competence with regard to perceiving student errors – or, more generally, to evaluate the correctness of students' answers – has already been measured in various empirical projects and studies.

Within the framework of the Mathematical Knowledge for Teaching (MKT) and Mathematics Teaching and Learning to Teach (MTLT) projects, Hill, Ball, and Schilling (2008) described the identification of common student errors and the provision of explanations as one of the major parts of teachers' mathematical knowledge for teaching. Ball, Thames, and Phelps (2008) summarized the perception of errors under the knowledge facets *Specialized Content Knowledge (SCK)* and *Knowledge of Content and Students (KCS)*, pointing out that error perception has two bases – mathematical knowledge and the knowledge of students and their relation to school mathematics. The latter kind of knowledge was described as the province of teachers, in contrast to mathematical knowledge, which resided in the province of professional mathematicians. Professional mathematicians were expected to be less familiar with common student errors. Further analyses could confirm that mathematicians were less likely to answer items correctly based on knowledge of students than content knowledge items (Hill, Dean and Goffney 2007). Ball, Thames and Phelps (2008) emphasized in their model that error analyses under time restriction are part of the characteristic tasks of teachers and clearly delineate their work differently from other professions, especially from the work of mathematicians.

The COACTIV study analyzed teachers' evaluations of students' solutions to mathematical tasks as either "correct" or "incorrect" (Binder et al. 2018) by measuring the teachers' reaction time and setting these results in relation to the number of student responses they assessed correctly. This construct is described as the *fast evaluation of elementary mathematical statements*. In contrast to the MKT project previously discussed, which focused on primary teachers, COACTIV evaluated the professional knowledge of secondary teachers. As COACTIV had in addition measured mathematical content knowledge and mathematics pedagogical content knowledge, the authors were able to estimate and examine the relations between the results from the test of reaction times and subjects' scores on these knowledge facets. The study showed a strong correlation between the fast evaluation of mathematical statements and mathematical content knowledge by the teachers as well as their mathematics pedagogical content knowledge (Krauss and Brunner 2011).

Like the MKT project, COACTIV analyzed the validity of its instruments and found – in contrast to the MKT project – that mathematicians solved the items significantly better, which supports the strong relation of the perception of student errors to mathematical content knowledge (Krauss, Baumert and Blum 2008).

The test of student-error perception that is presented in this paper, carried out as a separate test component within the TEDS-FU study, limited the time it took study participants to perceive student errors or assess the quality of students' statements. To bring the test closer to actual teaching environment, common student errors – ones that are often encountered in everyday teaching – were used as test items. Furthermore, the test was time limited in such a way that would make any recalculation by the participants impossible (Bühner 2011, p. 21). The number of correct answers was used as a performance indicator, and items not answered within the given time period were classified as wrong. The items used in the test for fast perception of student errors are of varying complexity (Pankow et al., 2016).

The present study examines the relations among mathematical content knowledge, mathematics pedagogical content knowledge, and fast perception of student errors using data from the TEDS-FU study. TEDS-FU measured the achievement of a subsample of early career (fourth-year) lower-secondary mathematics teachers in Germany who had participated in the TEDS-M study at the end of their teacher education. The data reveal a significant correlation of medium effect size between MCK and the fast perception of student errors ($r=.46$, $p<.001$; Blömeke et al. 2014, p. 528) and a significant correlation of weak effect size between MPCK and the error perception ($r=.28$).

Further studies of TEDS-FU data confirm the strong relation of mathematics pedagogical content knowledge and mathematical content knowledge with fast perception of student errors (Blömeke et al. 2016).

The dependence of error-perception speed on age has not yet been discussed in these studies, although many psychological studies point out that increasing age may bring with it a decrease in the speed with which many processing operations are executed, and that this reduction in speed may lead to disturbances in cognitive functions (Salthouse 1996). The decrease in sensory abilities in connection with the decrease in the ability for selective

attention and multi-tasking are especially well-documented, particularly in research on aging and driving (Madden 2007, Schlag 1993). This age dependency may play a role in the measurement of fast perception of student errors in TEDS-FU, where the speed of visual perception and physical reaction were emphasized.

Although the construct of fast perception of student errors was already described (Pankow et al. 2016), until now no external validation study concerning the test and its interpretation has been provided. Currently, only analyses of the internal structure of the whole set of tests applied within TEDS-FU exist, in which the relationships between the test components and its conformity to the construct are evaluated (Blömeke et al. 2016). In the following, we describe the results of external validation studies.

2.4. Approaches for validation

In educational research it is necessary to provide evidence for the quality of the instruments used in studies. Validity as a fundamental quality criterion “refers to the degree to which evidence and theory support the interpretation of tests scores for proposed uses of tests. Validity is, therefore, the most fundamental consideration in developing tests and evaluating tests” (AERA, APA, and NCME 2014, p. 11). In comparison to other quality criteria (objectivity and reliability) the examination of validity is more complex because standardized, routine procedures for supporting validity do not exist.

Classically, four main types of validity are distinguished: content or internal validity, criterion or external validity, statistical validity, and construct validity (Bortz and Döring 2006). Internationally, the concept of validity has changed over the years as the field of educational and psychological measurement has developed, and so far, there is no general agreement with regard to what exactly validity is and how it should be measured (e.g. Kane 2001; Lissitz and Samuelsen 2007; for a recent overview see Newton and Shaw 2016). One common view within contemporary discussions of validity is that in addition to the traditional sources of validity, such as evidence based on test content and expert opinions (traditionally called *content validity*) and evidence based on relations to other variables (an extension of the traditional *criterion validation*), response processes and consequences are also to be accepted as important sources of validity evidence that should be included in the validation process. Support for this view appears in the most recent edition of the *Standards for Educational and Psychological Testing* jointly published by the American Educational Research Association, the American Psychological Association, and the National Council on Measurement in Education (2014). Overall, these standards emphasize that “these sources of evidence may illuminate different aspects of validity, but they do not represent distinct types of validity. Validity is a unitary concept. It is the degree to which all the accumulated evidence supports the intended interpretation of test scores for the proposed use” (p. 13-14). Furthermore, the published standards emphasize the need to integrate various sources of validity evidence: “A sound validity argument integrates various strands of evidence into a coherent account of the degree to which existing evidence and theory support the intended interpretation of test scores for specific uses” (AERA, APA and NCME 2014, p. 21).

Following these standards, we examine evidence for test content as the most important source of validity and evidence for the relations to external variables. Content validity can be obtained, according to the standards,

. . . from an analysis of the relationship between the content of a test and the construct it is intended to measure. . . . The content specification carefully describes the content in detail, often with a classification of areas of content and types of items. Evidence based on test content can include logical or empirical analyses of the adequacy with which the test content represents the content domain and of the relevance of the content domain to the proposed interpretation of test scores. Evidence based on content can also come from expert judgements of the relationship between parts of the test and the construct. (AERA, APA and NCME 2014, p. 14)

Furthermore, the standards mention that evidence about content “can be used, in part, to address questions about differences in the meaning or the interpretation of test scores across relevant sub-groups of test takers” (AERA, APA and NCME 2014, p. 15). Analyses are therefore needed that explore the extent to which *construct underrepresentation* – i.e., the “degree to which a test fails to capture important aspects of the construct” (p. 12) – or *construct-irrelevance* – i.e., “the degree to which test scores are affected by processes that are extraneous to the test’s intended purpose” (AERA, APA and NCME 2014, p. 12) – may give an advantage to subgroups of test takers. This new kind of evidence for validity covers the traditionally distinguished *content validity* and *construct validity*.

A particular variation for providing evidence for construct validity is the *technique of related groups* (Bortz and Döring, 2006, p. 201). The selection characteristic used in this technique is the affiliation to composed groups, upon which the same test is carried out in order to measure different specifications of the construct. In other words, by deliberately constructing underrepresentation, it would become possible to study whether specific groups of test takers have an advantage and perform better or have a disadvantage and therefore perform poorer. In this study the related groups are selected by their empirically assumed expertise so that it will be relatively easy to immediately identify errors in mathematics. The test results should differ between participants who are able to perceive student errors under time limitation and those who are not (yet) able to do so. The group with a high qualification in MCK and MPCK should perform better in the test than those without any expertise in the area of MCK and MPCK. This particular technique has already been used in various studies in mathematics education, which will now be briefly described.

The COACTIV study empirically differentiated between content knowledge (CK) and pedagogical content knowledge (PCK) of mathematics teachers in Germany. To test the validity of the CK and PCK measurements in that study, researchers used contrast groups compared the results of COACTIV subjects with those of university students majoring in mathematics, biology and chemistry teachers, senior high school students, and mathematics teacher candidates at the end of their university education (Krauss et al. 2008). The mathematical content knowledge of all groups measured was at the levels hypothesized with mathematics majors outperforming all other groups and high school students as well as biology and chemistry teachers performing the worst. In contrast, the data did not provide

sufficient evidence for PCK as a distinct knowledge category of (future) mathematics teachers because the mathematics majors did well also on this test component (Krauss et al. 2008). In the empirical approach developed by Ball, Thames and Phelps (2008), contrast groups were used to provide evidence for the construct validity of a test of the professional knowledge of primary teachers, non-teachers of similar professions (e.g., nurses), and professional mathematicians (Hill, Dean and Goffney 2007). Participants were studied in cognitively oriented interviews to capture subject matter knowledge (SCK) and knowledge of content and students (KCS). In the interviews, 40 percent of the participants across all groups solved the items referring to KCS by using mathematical argumentation, which revealed that the student-focused material in those items was dominated by the underlying mathematical content; as a result, the items could not be considered as good measures of KCS (Ball et al. 2008).

Charalambous (2016) evaluated the knowledge of preservice elementary teachers, inservice elementary teachers and university students with strong mathematical background. The three groups performed all well on the pure mathematics content knowledge test while the results about mathematics pedagogical content knowledge were much in line with Krauss et al. (2008). Contrary to expectation, no significant differences could be identified between the two teacher groups and the university students in mathematically intensive departments. According to these results, MKT is measured too strongly in a holistic way and does not adequately reflect group differences. Charalambous (2016) concluded that more studies are needed to determine how teachers apply their specific professional knowledge and which role teachers' experience plays in developing teacher competence.

In addition, relations to other variables external to the test will be examined. "External variables may include measures of some criteria that the test is expected to predict, as well as relationships to other tests hypothesized to measure the same constructs, and tests measuring related or different constructs" (AERA, APA and NCME 2014, p. 16). Relations between test scores that aim to evaluate the same or similar constructs provide convergent evidence, whereas relations between test scores intending to measure different constructs deliberately offer discriminant evidence. A special case of relations are *test-criterion relationships*, which measure "how accurately the test scores predict criterion performance" (AERA, APA and NCME 2014, p. 17) and are called in classical terminology *criterion validity*.

In this study, evidence for criterion validity will be provided using the external criterion of the GPA of the university entrance qualification (so-called Abitur). This criterion has already been used in various studies in mathematics education, also in our own research in the TEDS-FU context. Blömeke et al. (2014) provided evidence that the university entrance qualification predicted fast perception of student errors mediated by the lower-secondary teachers' knowledge of mathematics content and mathematics pedagogical content.

3. Research aim and hypotheses

Using techniques based on the above described approaches, this study aims to provide evidence for the construct validity of the interpretations of the instrument used to measure fast

perception of student errors. Specifically, the validity assessment will be based on differences in the performance of contrast groups of test takers.

To achieve this goal, five contrast groups with varying expertise in mathematical knowledge, mathematics pedagogical content knowledge, teaching experience, and reaction time (based on age) were selected: future mathematics teachers in their master study (Sample 1: future teachers), teachers with four years of experience (Sample 2: early career teachers), teachers with more than four years of teaching experience (Sample 3: experienced teachers), university students studying mathematics as their degree program (Sample 4: university mathematics majors), and high school students (Sample 5: high school students).

This study will investigate the following hypotheses:

Hypothesis 1: Among all five contrast groups, the group of early career teachers will perform best in the test of fast perception of student errors. This hypothesis is based on three assumptions: early career teachers' experience in the classroom will enable them to perceive student errors more quickly than groups with less (future teachers) or without any teaching experience (mathematics majors and high school students), while the mathematical content knowledge members of this group gained during their teacher training is more recent – and therefore easier to recall – than that of experienced teachers, and their younger age allows them to react more quickly than this group.

Hypothesis 2: Among all five contrast groups, the group of high school students will perform worst in the test of fast perception of student errors due to lack of any teaching experience with these, and they will rank lowest in terms of both mathematical content knowledge and mathematics pedagogical content knowledge, as their training in both facets is the least developed of all participating groups.

Hypothesis 3: The three groups of future teachers, experienced teachers and mathematics majors will perform between the groups of early career teachers and high school students regarding fast perception of student errors.

Hypothesis 4: The results of the test of fast perception of student error can be predicted using students' GPA results from their university entrance qualification, which is known to be, in general, a strong indicator for academic success. This hypothesis is based on the results of previous studies, such as TEDS-M and TEDS-FU, that have shown that the grade of the Abitur turned out to be a valid predictor of academic success and fast perception of student error mediated by MCK and MPCK.

4. Design of the Study and Methodology

This section describes the instrument used to measure the fast perception of student errors, which was implemented as part of the national TEDS-FU study, as well as the participants of the five contrast samples.

4.1. Description of the test instrument

The fast perception of student errors is part of the range of facets that constitute teacher competence. The test instrument used to measure the fast perception of student errors consists

of 16 items, all of which are structured in the same way. This uniform presentation helps ensure that the instrument takes a uniform approach to the assessment of the different types of student errors and, thus, avoids methods bias.

The instrument was implemented online, and the tests took place at a self-chosen location with an internet connection (mainly at home). Since the participants were spread out across Germany, the online tool proved to be the ideal instrument for the test, since it allowed a widely disbursed sample of participants to complete the test in their own familiar environments, thus minimizing any distractions.

All student errors were derived from mathematical topics that constitute main themes of lower-secondary-level mathematics curriculum in Germany. Before the study began, participants were asked to think about their experience dealing with typical student errors, then an introductory item was shown to allow the participants to become acquainted with the test format. The test contained as specific feature an *anticipating phase*, i.e., before each student error was shown, the topic area from which the student error would come was shown to the participants (e.g., ‘addition of fractions’ in the case of Fig. 3, Item 1). The participants were asked to think about possible student errors coming from this mathematical topic. This phase was not strongly time-restricted – the participants could think as long as they considered appropriate, up to five minutes. This test design was developed to come closer to the reality of a classroom, where teachers know the topics they are teaching and have probably thought about common student errors or misconceptions in advance.

Once the participants decided to proceed with each item, they were shown three different student answers, of which one presented a typical student error that needed to be identified, while the other two answers were correct. The test was implemented in such a way that the participants were required to answer within four seconds. The maximum length of four seconds per item was determined by a group of 15 experts consisting of researchers from mathematics education and experienced teachers. The purpose of the time restriction was to prevent the participants from calculating the items in the given time, as this was not the purpose of the assessment; rather, the test was intended to measure the time it took to perceive the errors. If the test taker exceeded the time given, the response was considered incorrect. Practically, the test taker had to place one of three fingers of the left hand on each of the keys “s”, “d,” and “f” on a computer keyboard; these letters corresponded with the three given students answers, and the key indicating the erroneous student response had to be entered within the given time. Guess probability was 33 percent with each item.

For illustration purposes, two items of the test are displayed below (see Figure 3). The first item addresses the mathematical extension of addition from integers to rational numbers. According to Padberg (2009), this is an area of difficulty for many students that will continue to reappear throughout the course of students’ school careers. Wartha (2007) points out that nearly half of the errors made by students attempting to add two fractions are based on the approach of adding numerators and denominators separately. Additionally, students often develop their own set of rules when faced with the problem of adding two fractions (Padberg 2009).

The second item is less researched empirically, but well-known to practitioners. In German mathematics teaching, quadratic equations are usually solved with a specific formula after having standardized the quadratic equation to its normal form. The typical student error results while applying this formula $x_{1,2} = -\frac{p}{2} \pm \sqrt{\left(\frac{p}{2}\right)^2 - q}$. In addition to the probably missing standardization (Ritter and Voß 2015), the plus/minus sign ahead of the “q” is of importance – double-sign issues are likely to be encountered there, as shown in the example. Another difficulty is the p-q formula itself; it is, like other formulae and mathematical strategies, often applied without referring to its meaning (Allmendinger et al. 2013)

Nr.	Announcement	„s“	„d“	„f“
1	Fractions: Addition of two fractions	$\frac{1}{2} + \frac{1}{4} = \frac{3}{4}$	$\frac{2}{7} + \frac{3}{7} = \frac{5}{14}$	$\frac{1}{3} + \frac{2}{3} = 1$
4	Application of the p-q formula for quadratic equation $x^2 + 8x - 4 = 0$	$x^2 + 8x - 4 = 0$		
		$x_{1,2} = -4 \pm \sqrt{16 + 4}$	$x_{1,2} = -4 \pm \sqrt{16 - (-4)}$	$x_{1,2} = -4 \pm \sqrt{16 - 4}$

Fig. 3: Example items of “addition of two fractions” and “application of the solution formula for quadratic equations”

In sum, a variety of potential errors could be anticipated by the participant. These errors can be traced back to either the usage of the wrong concept by the student when adding fractions, or to the wrong application of the solution formula of quadratic equations, especially concerning the usage of the algebraic sign.

4.2. Samples

As already mentioned within the discussion of this study’s research goals, five different groups of participants were included in the validation study. Three groups were composed of teachers who were in different stages of their careers (see Table 1, Samples 1–3) and therefore had different levels of teaching experience. Two more groups were made up of students: one of university mathematics majors, the other, high school students. The two student groups were chosen to serve as contrast groups to the group of early career teachers included in the original TEDS-FU study. The university mathematics majors (Sample 4) are assumed to represent the highest accumulated level of MCK, whereas the senior high school students (Sample 5) are assumed to represent the lowest accumulated level of MCK. All three groups of teachers were assumed to fall somewhere between these two contrast sample groups.

This sample design was based on research in teacher expertise (Li and Kaiser, 2011), and groups were chosen to reflect the development of teachers’ expertise from the start of their initial training to later stages in their career. Furthermore, the increase in reaction time may reflect the increasing age of participants.

	Mathematician			Non-Mathematician
Teacher	Sample 1: Future teachers (n=42)	Sample 2: Early career teachers from TEDS-FU (n=137)	Sample 3: Experienced teachers from TEDS-Instruct (n=113)	/
Non-	Sample 4: University mathematics majors (n=151)			Sample 5: Senior

teacher		high school students (n=62)
---------	--	--------------------------------

Table 1: Samples and their sizes

In the following we will describe the various samples (Table 2):

Sample 1: The participating future teachers were in the first phase of their teaching career at the University of Hamburg. They had already completed their bachelor's degree for teaching at the primary or secondary level. All participants (n=42) were in an early stage of completing their master's degree, and the average age among them was 24 years (min.: 21; max.: 36; SD: 0.4). Males made up 11 percent of participants in this sample; 89 percent were female. They had completed their university entrance qualification (Abitur) with an average of 2.1 (SD=0.4), which is above the average of the North German Federal States, from where most students are originating, which is currently 2.5; the highest mark is 1.0.

Sample 2: The participants of TEDS-FU had originally participated in TEDS-M at the end of their German teacher education studies and had agreed to take part in a follow-up study 3.5 years later. They were spread all over Germany. Of the 171 data sets provided, 137 were valid for this test component. The participants' teaching experience ranged from three to four years, excluding their teaching traineeship. Their mean age was 32 years (SD=5.9). About two thirds of the sample were female (63%); all participants had completed their university entrance qualification with an average of 2.1 (SD=0.6).

Sample 3: These participants, experienced teachers from the German Federal State of Hamburg, took part in TEDS-Instruct, another follow-up study of TEDS-M. 113 participant data sets were deemed valid for use in the analysis. The participants indicated a broad range of teaching experience. Just over half of the participants in the sample were female (50.4%); the mean age was 39 years (SD=10.4). Participants in this sample had completed their university entrance qualification with an average of 1.71 (SD=0.5).

Sample 4: These participants were studying either at the University Hamburg or the University of Bremen. They were at the end of their bachelor study and aiming for a degree in mathematics. Their participation was voluntary and by the invitation of the university lecturer, who was asked to help recruit participants for the study. This group had an average age of 21 years (SD=4.06); 26 percent of its members were female, and they had completed its university entrance qualification with an average of 1.87 (SD=0.4). During their bachelor program, the students had participated in 4.7 (SD=3.1) mathematics courses.

Sample 5: The high school students were tested in two mathematics courses at a comprehensive school in Hamburg. The participants ranged in age from 17 to 21 years. This sample was 37 percent female, and all students in the class participated in the survey.

Sample	N	Age (years)	Gender	University entrance qualification
Sample 1: Future teachers	42	24 (SD= 0.4).	89% female	2.1 (SD=0.4),
Sample 2: Early career teachers (data from TEDS-FU)	137	32 (SD=5.9).	63% female	2.1 (SD=0.6)

Sample 3: Experienced teachers (data from TEDS-Instruct)	113	39 (SD=10.4).	50% female	1.71 (SD=0.5).
Sample 4: University mathematics majors	151	21 (SD=4.06)	26% female.	1.87 (SD=0.4)
Sample 5: Senior high school students	62	17 – 21 (estimated)	37% female	/

Table 2: Demographic details of the five sample groups

4.3. Methodology

In order to compare the test results across groups, a certain level of *measurement invariance* is a necessary requirement. Testing for the equivalency of the measurement ensures that the same construct is measured in every sub-sample. Measurement invariance is often tested by constraining parameters in a confirmatory factor analysis, evaluating the model fit, and comparing it with the fit of a less restrictive model. Followed the approach taken by Krauss et al. (2008) subscales were constructed that represent typical student errors in handling fractions, functions, or algorithmic or arithmetic problems (four items each). In all subscales, a participant was able to achieve a score between 0 and 4. These subscales were then used as indicators through the process of item-parceling (see Bandalos and Finney 2009). The indicators were used to conduct a series of confirmatory factor analyses with lavaan 0.5 series (Rosseel 2017) under R version 3.4.3. First, the same measurement model was applied to all groups – i.e., the construct was operationalized by the same items within every group. In this first step, the parameters (such as factor loadings and intercepts) were estimated freely. Second, the factor loadings were constrained, which results in so-called metric invariance. Metric invariance is required if one is to conclude that the same construct is being examined in all subpopulations (see Schwab & Helm 2015 for more information). In a third step, further parameters such as intercepts and residual variances may be constrained (scalar invariance). Scalar invariance ensures that all the items were measured on the same level. Hence, absolute comparisons on the latent variable between groups would be possible.

5. Results

A multi-group confirmatory factor analysis with subscales as indicators showed an almost perfect fit for both configural ($\chi^2 = 8.64$, $df = 10$, $p = .56$, CFI = 1.00, RMSEA = .00) and metric invariance ($\Delta\chi^2 = 16.95$, $\Delta df = 12$, $p = .77$). However, restricting the model further to scalar invariance resulted in a significantly worse model fit ($\Delta\chi^2 = 81.42$, $\Delta df = 12$, $p < .01$). Comparisons of mean group results have therefore be used with care whereas the prerequisites for a quasi-experimental design with non-equivalent control groups as applied in this study should be given.

The following explores the question of the degree to which the evidence for construct validity of the fast perception of student errors can be supported based on the performance analysis of different sample groups (see Table 3 and Fig. 4).

In contrast to the first hypothesis, the early career teachers did not perform best on the test of fast perception of student errors; they were coming in between the university mathematics majors and high school students. The Hypothesis 2 could be confirmed though, the group of high school students received the lowest results, with a median of 3 correct answers – lower than the probability of guessing (5 correct answers). Also hypothesis 3 was supported by the

data with respect to future and experienced mathematics teachers who came in between mathematics majors and high school students.

Sample no.	Sample	N	Median (min.; max.)
4	University mathematics majors	151	9 (0;16)
1	Future mathematics teachers	42	8 (4;13)
2	Early career teachers (TEDS-FU)	137	7 (0;14)
3	Experienced teachers (TEDS-Instruct)	113	6 (0;13)
5	Senior high school students	62	3 (0;9)

Table 3: Sample and their correct answers

A Kruskal-Wallis H test showed that there was a statistically significant difference in the average scores received on the timed test between the groups, $\chi^2(4) = 101.374$, $p = 0.000$, with a rank score of 313.62 for university mathematics majors, 260.98 for future mathematics teachers, 255.44 for early career teachers (TEDS-FU), 217.80 for experienced teachers (TEDS-Instruct) and 105.15 for senior high school students.

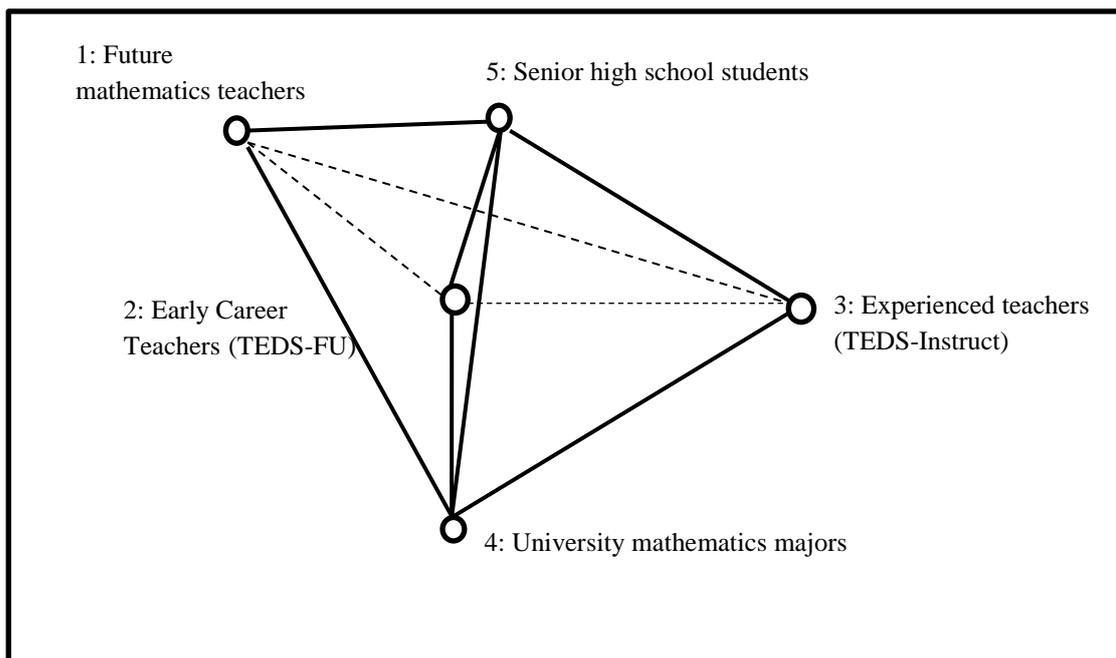


Fig. 4: Pairwise comparison of the different samples [dashed lines mean no significant difference; continuous lines indicate significant difference].

A pairwise comparison, shown in Figure 4, analyzes the average ranking of the groups and examines statistical significance. The displayed knots represent the groups, bold connections describe highly significant differences between the groups, and the dashed lines show that the connected groups do not differ significantly. These results show that the groups of teachers do not significantly differ from each other, whereas the two contrast groups – the senior high school students and the university mathematics majors – differ significantly from all other groups.

All participants in the original study and in the control groups were asked to disclose their age. Although 258 participants answered this question, none of the high school students did, so they were excluded from this analysis and its result. The results point to a negative correlation between success on the test and age, $r_s = -.227$ ($p = .000$), meaning that as the age of the test takers increase, test scores tend to decrease.

Concerning the fourth hypothesis – the ability to predict the speed of student-error perception based on university entrance qualification scores – the following results were achieved. Data on 359 participants show a small but highly significant relation between university entrance examination results and achievement in the test on the fast recognition of student errors, $r_s = -.14$ ($p < .001$). These results are in line with those of previous studies such as TEDS-M (Blömeke, Kaiser and Lehmann 2010) or TEDS-FU (Blömeke et al. 2014).

6. Summary and discussion of the main results

On average, the university mathematics majors received the highest scores in the test of fast perception of student errors, which is similar to, though not exactly the same as, the results in the COACTIV study. This result indicates the importance of sound mathematics content knowledge by mathematics teachers for perceiving student errors.

The results found among the various groups of teachers were significantly better than the results of the high school students, which supports the high relevance of mathematical content knowledge and of mathematics pedagogical knowledge for the fast perception of student errors; the high school students were expected to possess low knowledge in both domains.

Results among the various groups of practicing and future teachers were not very distinct from one another, which implies that teaching experience in our case does not play a strong role in determining the speed of student-error perception. One reason for this may be that the student errors in the test instrument were not embedded in the type of pedagogically rich situation that occurs in the classroom; one in which the teachers would have anticipated possible student errors already in their lesson preparation and where the knowledge of strengths and weaknesses of the individual student often plays a significant role. Outside this pedagogical environment, mathematical content knowledge seems to play the most relevant role.

University entrance qualification results could be used as an external variable to predict the test results of participants. Except for the high school students, who had not yet passed this examination at the test time, evidence for the validity of the instrument could be provided by significant correlations between academic success at the end of school and the results in the test of fast perception of student errors for every sample in the study. The results for fast perception of student errors could confirm the assumed ranking of the groups according to their expertise and therefore confirm the evidence for validity based on relations to an external variable, namely university entrance qualification exam results, as extension of the criterion validity of the fast perception of student errors.

As to limitations of the study, it can be seen that only parts of the error perception cycle of Heinrichs and Kaiser (2018) were tested. Neither reasons for the perceived student error nor

possible pedagogical measures were included in the test, although they belong to a comprehensive understanding of diagnostic competence as a sub-facet of mathematics pedagogical competence. To include these more situated aspects of error perception, more contextualized items would be necessary – such as those used by, for example, Heinrichs and Kaiser (2018). However, doing so could affect the clearness of the results seen here, as various other knowledge facets and indeterminable influences would come into play. Finally, it has to be considered that “it is commonly observed that the validation process never ends, as there is always additional information that can be gathered to more fully understand a test and the inference that can be drawn from it” (AERA, APA, and NCME 2014, p. 21-22).

As another limitation we have to point out that scalar measurement invariance did not exist which limits the possibility to carry out group mean comparisons. However, a contrast group approach can be regarded as a type of quasi-experimental design with non-equivalent groups where metric invariance should be sufficient.

Overall, the fast perception of student errors in pedagogical situations creates important opportunities to assess the level of students’ knowledge and to help them depart from this level by focusing, if necessary, on students’ misconceptions or other learning difficulties. Therefore, being able to perceive a student’s error quickly is a key competence for quality-oriented instruction and needs to be stressed in teacher education, by fostering not only mathematics pedagogical content knowledge but also mathematical content knowledge as bases for this perception.

References

- Allmendinger, H., Lengnink, K., Vohns, A., & Wickel, G. (2013). *Mathematik verständlich unterrichten: Perspektiven für Unterricht und Lehrerbildung*. Wiesbaden: Springer.
- Altmann, A. F., & Nückles, M. (2017). Empirische Studien zu Qualitätsindikatoren für den diagnostischen Prozess. In A. Südkamp & A.-K. Praetorius (Eds.), *Diagnostische Kompetenz von Lehrkräften. Theoretische und methodische Weiterentwicklungen* (pp. 142–149). Münster: Waxmann.
- Ball, D., Thames, M. H., & Phelps, G. (2008). Content Knowledge for Teaching: What Makes It Special? *Journal of Teacher Education*, 59(5), 389–407.
- Bandalos, D. L. & Finney, S. J. (2009). Item Parceling Issues in Structural Equation Modeling. In G.A. Marcoulides, & R.E. Schumacker (Eds.), *New Developments and Techniques in Structural Equation Modeling* (pp. 269-296). Mahwah: Erlbaum.
- Binder K., Krauss S., Hilbert S., Brunner M., Anders Y., Kunter M. (2018) Diagnostic Skills of Mathematics Teachers in the COACTIV Study. In T. Leuders, K. Philipp K., & J. Leuders (Eds.), *Diagnostic Competence of Mathematics Teachers* (pp. 33-53). Mathematics Teacher Education. Cham: Springer.
- Blömeke, S., Kaiser, G., & Lehmann, R. (2008) (Eds.). *Professionelle Kompetenz angehender Lehrerinnen und Lehrer. Wissen, Überzeugungen und Lerngelegenheiten deutscher Mathematik-Studierender und -referendare – Erste Ergebnisse zur Wirksamkeit der Lehrerausbildung*. Münster: Waxmann.

- Blömeke, S., Kaiser, G., & Lehmann, R. (2010) (Eds.). *TEDS-M 2008 – Professionelle Kompetenz und Lerngelegenheiten angehender Mathematik-Lehrkräfte für die Sekundarstufe I im internationalen Vergleich*. Münster: Waxmann.
- Blömeke, S., Gustafsson, J.-E., & Shavelson, R. J. (2015). Beyond Dichotomies. Competence Viewed as a Continuum. *Zeitschrift für Psychologie*, 223(1), 3–13.
- Blömeke, S., König, J., Busse, A., Suhl, U., Benthien, J., Döhrmann, M., & Kaiser, G. (2014). Von der Lehrerausbildung in den Beruf – Fachbezogenes Wissen als Voraussetzung für Wahrnehmung, Interpretation und Handeln im Unterricht. *Zeitschrift für Erziehungswissenschaft*, 17(3), 509–542.
- Blömeke, S., Busse, A., Kaiser, G., König, J., & Suhl, U. (2016). The relation between content-specific and general teacher knowledge and skills. *Teaching and Teacher Education*, 56 (May), 35-46
- Bortz, J., & Döring, N. (2006). *Forschungsmethoden und Evaluation: Für Human- und Sozialwissenschaftler* (4. ed.) Heidelberg: Springer.
- Bühner, M. (2011). *Einführung in die Test- und Fragebogenkonstruktion*. München: Pearson Deutschland.
- Charalambous, C. Y. (2016). Investigating the Knowledge Needed for Teaching Mathematics. *Journal of Teacher Education*, 67(3), 220–237.
- Helmke, A. (2015). *Unterrichtsqualität und Lehrerprofessionalität. Diagnose, Evaluation und Verbesserung des Unterrichts* (6th rev. ed.). Seelze: Klett-Kallmeyer.
- Heinrichs H., & Kaiser G. (2018) Diagnostic Competence for Dealing with Students' Errors: Fostering Diagnostic Competence in Error Situations. In T. Leuders, K. Philipp, & J. Leuders (Eds.) *Diagnostic Competence of Mathematics Teachers* (pp. 79-94). Cham: Springer.
- Heinze, A. (2004). Zum Umgang mit Fehlern im Unterrichtsgespräch der Sekundarstufe I. *Journal für Mathematik-Didaktik*, 25(3-4), 221–244.
- Herppich, S., Altmann, A. F., Wittwer, J., & Nückles, M. (2017). Förderung von Instruktionsstrategien zum verbesserten Diagnostizieren im Eins-zu-Eins-Tutoring. In A. Südkamp, & A.-K. Praetorius (Eds.), *Diagnostische Kompetenz von Lehrkräften. Theoretische und methodische Weiterentwicklungen* (pp. 203–208). Münster: Waxmann.
- Hill, H. C., Ball, D., & Schilling, S. (2008). Unpacking Pedagogical Content Knowledge: Conceptualizing and Measuring Teachers' Topic-Specific Knowledge of Students. *Journal for Research in Mathematics Education*, 39(4), 372–400.
- Hill, H. C., Dean, C., & Goffney, I. M. (2007). Assessing Elemental and Structural Validity: Data from Teachers, Non-teachers, and Mathematicians. *Measurement: Interdisciplinary Research & Perspective*, 5(2-3), 81–92.
- Hoth, J., Döhrmann, M., Kaiser, G., Busse, A., König, J., & Blömeke, S. (2016). Diagnostic competence of primary school mathematics teachers during classroom situations. *ZDM Mathematics Education*, 48(1), 41–53.
- Kane, M. T. (2001). Current concerns in validity theory. *Journal of educational measurement*, 38(4), 319-342.

- Kleickmann, T., Richter, D., Kunter, M., Elsner, J., Besser, M., Krauss, S., & Baumert, J. (2013). Teachers' Content Knowledge and Pedagogical Content Knowledge: The Role of Structural Differences in Teacher Education. *Journal of Teacher Education*, 64(1), 90-106.,
- Krauss, S., Baumert, J., & Blum, W. (2008). Secondary mathematics teachers' pedagogical content knowledge and content knowledge: Validation of the COACTIV constructs. *ZDM Mathematics Education*, 40(5), 873–892.
- Krauss, S., Blum, W., Brunner, M., Neubrand, M., Baumert, J., Kunter, M., Besser, M. & Elsner, J. (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. Results from the COACTIV Project* (pp. 147-174). New York: 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., & Brunner, M. (2011). Schnelles Beurteilen von Schülerantworten: Ein Reaktionszeittest für Mathematiklehrer/innen. *Journal für Mathematik-Didaktik*, 32(2), 233–251.
- Leuders, T. (2001). *Qualität im Mathematikunterricht in der Sekundarstufe I und II*. Berlin: Cornelsen Scriptor.
- Lindmeier, A. M., Heinze, A., & Reiss, K. (2013). Eine Machbarkeitsstudie zur Operationalisierung aktionsbezogener Kompetenz von Mathematiklehrkräften mit videobasierten Maßen. *Journal für Mathematik-Didaktik*, 34(1), 99–119.
- Lissitz, R. W., & Samuelsen, K. (2007). A suggested change in terminology and emphasis regarding validity and education. *Educational Researcher*, 36(8), 437-448.
- Madden, D.J. (2007). Ageing and Visual Attention. *Current Directions in Psychological Science*, 16, 2, 70-74.
- Newton, P., & Shaw, S. (2016) Disagreement over the best way to use the word 'validity' and options for reaching consensus. *Assessment in Education: Principles, Policy & Practice*, 23(2), 178-197
- Oser, F., Hascher, T., & Spychiger, M. (1999). Lernen aus Fehlern: Zur Psychologie des "negativen" Wissens. In W. Althof (Ed.), *Fehlerwelten. Vom Fehlermachen und Lernen aus Fehlern* (pp. 11–41). Wiesbaden: VS Verlag für Sozialwissenschaften.
- Padberg, F. (2009). *Didaktik der Bruchrechnung*. Heidelberg: Springer.
- Pankow, L., Kaiser, G., Busse, A., König, J., Blömeke, S., Hoth, J., & Döhrmann, M. (2016). Early Career Teachers' ability to focus on typical students errors in relation to the complexity of a mathematical topic. *ZDM Mathematics Education*, 48(1-2), 55–67.
- Radatz, H. (1980a). *Fehleranalysen im Mathematikunterricht*. Wiesbaden: Vieweg+Teubner Verlag.
- Radatz, H. (1980b). Students' Errors in the Mathematical Learning Process: A Survey. *For the Learning of Mathematics*, 1, 16–20.
- Reisman, F. K. (1976). *A guide to the diagnostic teaching of arithmetic*. Columbus: Merrill.

- Ritter, S., & Voß, U. (2015). *Erfolgreich Starten ins Ingenieurstudium: Grundlagen der Mathematik anwendungsorientiert erklärt*. Berlin: Springer Vieweg.
- Rosseel, Y. (2017). *Package "lavaan": Latent variable analysis. Comprehensive R Archive Network*. Retrieved from <http://lavaan.org>
- Salthouse, T. A. (1996). The processing-speed theory of adult age differences in cognition. *Psychological Review*, 103(3), 403-428.
- Schlag, B. (1993). Elderly drivers in Germany - Fitness and driving behavior. *Accident Analysis and Prevention*, 25 (1), 47-55.
- Schwab, S., & Helm, C. (2015). Überprüfung von Messinvarianz mittels CFA und DIF-Analysen. *Empirische Sonderpädagogik*, 7 (2015) 3, 175-193
- Schoy-Lutz, M. (2005). *Fehlerkultur im Mathematikunterricht: Theoretische Grundlegung und evaluierte unterrichtspraktische Erprobung anhand der Unterrichtseinheit "Einführung in die Satzgruppe des Pythagoras"*. Hildesheim: Franzbecker.
- Seidel, T., & Prenzel, M. (2003). *Mit Fehlern umgehen - Zum Lernen motivieren. Praxis der Naturwissenschaften - Physik*, 51(1), 30–34.
- Shulman, L. S. (1986). Those Who Understand: Knowledge Growth in Teaching. *Educational Researcher*, 15(2), 4–14.
- Südkamp, A., & Praetorius, A.-K. (Eds.) (2017). *Diagnostische Kompetenz von Lehrkräften: Theoretische und methodische Weiterentwicklungen*. Münster: Waxmann.
- Swan, M. (2004). Making sense of mathematics. In I. Thompson (Ed.), *Enhancing primary mathematics teaching and learning* (pp. 111–124). Maidenhead: Open University Press.
- Türling, J. M. (2014). *Die professionelle Fehlerkompetenz von (angehenden) Lehrkräften: Eine empirische Untersuchung im Rechnungswesenunterricht*. Wiesbaden: Springer.
- Wahl, D., Weinert, F. E., & Huber, G. L. (1984). *Psychologie für die Schulpraxis: Ein handlungsorientiertes Lehrbuch für Lehrer*. München: Kösel.