Understanding the Formation and Improving the Accuracy of Teacher Judgment

Inauguraldissertation

zur Erlangung des Doktorgrades der Philosophie

im Fach Psychologie an der Universität Passau

vorgelegt von

Chunjie Zhu

April 2019

UNIVERSITÄT PASSAU

Philosophische Fakultät

Erstgutachter: Prof. Dr. Detlef Urhahne

Zweitgutachterin: Prof. Dr. Jutta Mägdefrau

ZUSAMMENFASSUNG

Die Forschung zu Lehrkrafturteilen hat in den letzten drei Jahrzehnten beträchtliche Fortschritte gemacht. Die Bedeutung des Lehrkrafturteils und die Variabilität in der Urteilsgenauigkeit erfordern eine eingehendere Untersuchung. Basierend auf der Überprüfung früherer Studien wurde ein systematischer analytischer Rahmen vorbereitet, der aus drei Hauptstudien besteht, um das Verständnis der Prozesse und Merkmale von Lehrkrafturteilen zu erweitern. In den drei vorgestellten Studien wurde insbesondere untersucht, wie Lehrkrafturteile durch verschiedene Schülermerkmale generiert werden, welche Möglichkeiten es gibt, die Urteilsgenauigkeit von Lehrkräften zu verbessern und ob die Urteilsgenauigkeit von Lehrkräften im Laufe der Zeit stabil bleiben kann.

In der ersten Studie wurde das Linsenmodell der Theorie der sozialen Beurteilung angewendet, um die Einschätzungen von Lehrpersonen über die Leistung von Schülerinnen und Schülern und ihre Strategien der Informationsverarbeitung besser zu verstehen. 260 Lehrkräfte aus sieben chinesischen Grundschulen wurden gebeten, aus sieben Informationsquellen Schülermerkmale auszuwählen und zu bewerten, anhand derer sie die Leistungen der Schüler beurteilen könnten. Die Lehrpersonen entwickelten eine klare Hierarchie der verwendeten Datenquellen. Die besten Informationen wurden aus den Fähigkeiten und Einstellungen der Schülerinnen und Schüler gewonnen und die am wenigsten wichtigen Informationen aus der sozialen Interaktion mit anderen sowie aus der Schüler-Demografie. Um genauere Einschätzungen zu treffen, sollten die Lehrkräfte über gültige Indikatoren für die Schülerleistung informiert werden.

Die zweite Studie zielte darauf ab, die Urteilsgenauigkeit von Lehrkräften und die Leistung der Schülerinnen und Schüler durch den Einsatz von Classroom-Response-Systemen („Clickern“) zu fördern. 20 Schulklassen mit 459 Schülerinnen und Schülern der sechsten Klasse und ihren Mathematiklehrkräften wurden für eine fünfwöchige quasi-experimentelle Interventionsstudie mit einem Pre- und Post-Test in drei Gruppen eingeteilt. Die Ergebnisse

zeigen, dass beide Ziele weitgehend erreicht werden konnten. Schülerinnen und Schüler der Clicker-Gruppe haben durch die Intervention mehr mathematisches Wissen erworben als Studenten der Tagebuch- und Kontrollgruppe. Die Lehrkrafturteile aller drei Gruppen wurden vom Pre- zum Post-Test genauer. Lehrpersonen, die Clicker verwendeten, beurteilten jedoch mit höchster Genauigkeit. Clicker können als wertvolles Werkzeug zur Verbesserung der Urteilsgenauigkeit von Lehrkräften empfohlen werden.

In der dritten Studie wurde die zeitliche Stabilität der Urteilsgenauigkeit der Lehrkräfte hinsichtlich Motivation, Emotion und Leistung der Schülerinnen und Schüler untersucht. Neun Klassen mit 326 Sechstklässlern einer chinesischen Grundschule und ihren Mathematiklehrpersonen nahmen an der Studie teil. Die Schüler arbeiteten an einem standardisierten Mathematik-Test und einem Selbstbeschreibungsfragebogen zu Motivation und Emotion. Die Lehrpersonen beurteilten die Motivation, Emotion und Leistung jedes einzelnen Schülers anhand einzelner Items. Das Lehrkrafturteil und die Eigenschaften der Schülerinnen und Schüler wurden innerhalb von vier Wochen zweimal gemessen. Die Ergebnisse zeigten, dass die Lehrkräfte in der Lage waren, die Schülerleistungen mit hoher Genauigkeit, die Motivation der Schülerinnen und Schüler mit mäßiger bis hoher Genauigkeit und die Emotion der Schülerinnen und Schüler meist mit geringer Genauigkeit zu bewerten. Die Urteilsgenauigkeit der Lehrpersonen war sehr stabil mit nur geringen Veränderungen an den verschiedenen Genauigkeitskomponenten. Es kann gefolgert werden, dass chinesische Grundschullehrkräfte in der Lage sind, zu verschiedenen Zeitpunkten faire Urteile über Schülerleistungen und der Motivation ihrer Schülerinnen und Schüler zu treffen. Die Emotionen der Schülerinnen und Schüler sind für Lehrpersonen jedoch schwer zu erfassen.

ABSTRACT

Research on teachers' judgment has made considerable progress in the last three decades. The significance of teacher judgment and the variability in judgment accuracy warrant deeper investigation. Based on a review of previous studies, a systematic analytical framework that consists of three main studies was prepared to broaden the understanding of the processes and features of teacher judgment. In particular, the three studies examine how teacher's judgment is generated from different types of student cues, what are the possible ways to improve teacher judgment accuracy, and whether teacher's judgment accuracy could remain stable over time.

In the first study, the lens model of social judgment theory was applied to better understand teachers' judgments of student achievement and their strategies of information processing. Two-hundred and sixty teachers from seven Chinese primary schools were asked to select and rank student cues from seven information sources that would help them to judge student achievement. Teachers developed a clear hierarchy of utilized data sources. The best information was rearing from student abilities and attitudes and the least important information from social interaction with others and student demographics. To make more accurate judgments, teachers should be informed about more valid indicators of student achievement.

The second study aimed to support teacher judgment accuracy and student achievement by the use of learner response systems ("clickers") in the classroom. Twenty school classes with 459 sixth-grade students and their mathematics teachers were divided into three groups for a quasi-experimental pre-post-test intervention study over five weeks. The results indicate that both objectives could be achieved to a large extent. Students of the clicker group gained more mathematical knowledge from the intervention than students of the diary and control group. Teacher judgments of all three groups were getting more accurate from

pre- to post-test. However, teachers using clickers judged with far the highest accuracy. Clickers can be recommended as a valuable tool for enhancing teacher judgment accuracy.

Temporal stability of teachers' judgment accuracy of students' motivation, emotion, and achievement was examined in the third study. Nine classes with 326 sixth-graders from a Chinese elementary school and their mathematics teachers took part in the study. Students worked on a standardized mathematics test and a self-description questionnaire for measuring students' motivation and emotion. Teachers judged each student's motivation, emotion and achievement by single items. The correspondence between teacher judgments and student characteristics was measured twice within a four-week period in order to determine the accuracy of teacher judgment over time. The results showed that teachers were able to assess student achievement with high accuracy, student motivation with moderate to high accuracy, and student emotion mostly with low accuracy. Teachers' judgment accuracy was highly stable with little changes on different accuracy components. It can be concluded that Chinese elementary school teachers are in a position to make fair judgments about student achievement and student motivation at different times. Student emotions, however, are hard to grasp for teachers.

Table of Contents

CHAPTER 1: GENERAL INTRODUCTION

1.1 The Meaning of Teacher Judgment

Judgment refers to the process of evaluation or categorizing a person or an object (Shavelson & Stern, 1981; Shulman & Elstein, 1975). One of the first attempts to conceptualize teacher judgment was from Varner (1923) who argued that teachers are continually required to estimate the traits of their students (Shavelson, 1983). In the study, he checked the ability of teachers to estimate students' intelligence and found that their judgments were inaccurate. He prepared the ground for a number of subsequent studies. In these, teacher judgment is defined as teachers' estimation of students' attributes (Hoge & Coladarci, 1989; Südkamp, Kaiser, & Möller, 2012; Zhu, Urhahne, & Rubie-Davies, 2018).

Teacher judgments are explored as one of the most important teachers' cognitive processes. Therefore, it is indispensable to distinguish teacher judgment from other teachers' cognitive behaviors, i.e., teacher expectations and teacher decisions. As mentioned, teacher judgments are teachers' estimates of students' current status. Teacher expectations are defined as teachers' inferences about future behavior or academic achievement of students according to their current status (Good, 1987). In contrast, teacher decisions are conscious selections of some specific actions (Heald, 1991). It was also found that many decisions made in the educational context are based on teachers' judgments (Glogger-Frey, Herppich, & Seidel, 2018; Heald, 1991; Hoge & Coladarci, 1989).

1.2 The Accuracy of Teacher Judgment

Since the beginning of the 21st century, there is a growing awareness concerning the importance of teachers' judgments about specific students' aspects and their judgment accuracy (Südkamp et al., 2012). Numerous empirical research studies have examined teachers' judgment accuracy of students' achievement, motivation, and emotions.

1.2.1 The Importance of Teacher Judgment Accuracy

Teacher judgment is of exceptional importance, it can have consequences for both the practice of teaching and the improvement of learning. Teacher judgment accuracy is often

necessary to exactly assess students or groups of students on aspects such as achievement, intelligence or learning difficulties. Accurate judgments assist teachers in fostering equal opportunities for all students in class (Paleczek, Seifert, & Gasteiger-Klicpera, 2017). They are helpful in identifying students with special needs and making further counseling decisions (Bates & Nettelbeck, 2001; Hoge & Coladarci, 1989).

Teacher judgments have the function of providing feedback to students and their parents (Alvidrez & Weinstein, 1999). This is especially crucial in stratified education systems, where students are assigned to different types of secondary schools mainly depending on teacher-assessed academic performance.

Finally, teachers' judgments can influence the expectations about students' ability and lead to teacher behaviors that stimulate students' motivation, emotion and achievement (Brophy & Good, 1970; Urhahne, 2015). Teacher judgments have the potential to create self-fulfilling prophecies (Babad, 1993; Jussim, 1989), which can crucially impact students' academic self-concepts and vocational careers (Südkamp et al., 2012). All this research illustrates the significance of teacher judgment accuracy and justifies a deeper discussion of the subject.

1.2.2 Measuring Teacher Judgment Accuracy

The accuracy of teacher judgment about students can be determined by three different components: rank, level, and differentiation component (Cronbach, 1955). The rank component indicates whether the teacher can rank students well with respect to certain characteristics. Therefore, class-wise calculated Pearson correlations between teacher judgments and student characteristics are Fisher-z-transformed. The mean Fisher-z value is transformed back into a Pearson correlation which represents the rank component (Helmke & Schrader, 1987). The level component shows whether the teacher can correctly judge the level of a class. The level component is given by the difference between teacher judgment and student characteristic. The differentiation component indicates whether the teacher correctly

assesses the heterogeneity of student characteristics in class. The differentiation component is the mean within-class variance of teacher judgments divided by the variance of the student characteristics (Helmke & Schrader, 1987).

The rank component is considered to be the most important indicator for determining the accuracy of teacher judgment (Madelaine & Wheldall, 2005). Like a correlation, it can vary between minus one and plus one, with positive values being the rule and negative values being the exception (Machts, Kaiser, Schmidt, & Möller, 2016; Südkamp et al., 2012). In order to make fair judgments, it is not only important that teachers accurately assess students' rank order. Students' level has to be considered as well. Thiede et al. (2018) speak of absolute accuracy as opposed to relative accuracy of the rank component. The level component is not limited to a specific range of values, but a value of zero is considered ideal as there are on average no differences between teacher judgments and student attribute. If misjudgments on the level component occur, students are judged either too positive (values greater than zero) or too negative (values less than zero). Research has also shown that underestimating students' achievement is associated with a large number of motivational and emotional deficits on part of the students (Urhahne, Chao, Luttenberger, Florineth, & Paechter, 2011). The differentiation component does not have the same meaning as the other two components since educational conclusions are much harder to draw. The differentiation component has an ideal value of one in case that the variability of teachers' judgments and students' characteristics are congruent to each other. Values above one indicate overestimation and values below one stand for underestimation of the variability of student characteristics in a class.

1.2.3 Teacher Judgment Accuracy of Student Achievement

In the existing studies, the most measured aspect of teachers' judgment accuracy is student achievement (Hoge & Coladarci, 1989; Südkamp et al., 2012). To determine the judgment accuracy of student achievement, teachers' estimations are usually compared with students' academic performance in a standardized test. On the one side, empirical findings on

the accuracy of teacher judgment on student performance were concluded to be on a moderate to strong level. For example, Hoge and Coladarci (1989) found a correlation of 0.66 between teacher judgment and student performance in a standardized test. Südkamp et al. (2012) in another meta-analysis reported a mean effect size of 0.63.

On the other side, teacher judgment accuracy varied significantly in different studies. Correlations in the recent meta-analysis were found to be ranging from r = -.03 to r = .84 (Südkamp et al., 2012). These findings indicated that there are apparent individual differences among teachers' judgment accuracy. Some teachers could predict their students' performance very well, whereas some others seem inexperienced and failed to judge students correctly.

1.2.4 Teacher Judgment Accuracy of Student Motivation and Emotions

In addition to student academic achievement, teacher judgment accuracy regarding student motivation and emotions has been the object of some studies (Givvin, Stipek, Salmon, & MacGyvers, 2001; Helm, Müller-Kalthoff, Mukowski, & Möller, 2018; Praetorius, Berner, Zeinz, Scheunpflug, & Dresel, 2013; Spinath, 2005; Urhahne et al., 2011; Urhahne, Timm, Zhu, & Tang, 2013). To make instructional decisions or provide feedback to parents, teachers are expected to know whether their students are self-confident and willing to make an effort, study with interest, or anxiously look forward to the upcoming exams. In turn, teachers' perceptions on students' learning motivation could have influence on students' emotion and knowledge acquisition (Zhu & Urhahne, 2014).

Compared with judgment accuracy of student achievement, teachers are found to have more difficulties to assess students' motivation and emotions (Karing, 2009; Karing, Dörfler, & Artelt, 2015; Spinath, 2005; Wright & Wiese, 1988; Urhahne et al., 2010; Zhu & Urhahne, 2014). For example, Spinath (2005) reported that teachers could judge students' academic self-concept (r = .39) with moderate accuracy and learning motivation (r = .20) as well as test anxiety (r = .15) with comparatively low accuracy.

The deviations between teachers' judgments and students' motivational-affective traits demonstrate the necessity of adjustment. It is therefore meaningful to conduct more research focusing not only on teachers' judgment level of students' motivation, but also on the development of their judgment accuracy.

1.3 The Base of Teacher Judgment

The variability of teacher judgment accuracy suggests exploring the reasoning behind teachers' judgments in order to explain the discrepancies. In another word, it is meaningful to consider what factors or information teachers actually use in order to form their judgments.

1.3.1 Modeling the Teacher Judgment Process

Teacher judgment is regarded as a cognitive process in a sophisticated context (Haigh, Ell, & Mackisack, 2013; Haigh & Ell, 2014) and a set of approaches were applied to analyze this process. According to Shavelson (1983), the formation of teacher judgment could be considered as classification, selection, and estimation and was described as follows:

> Teachers have available a large amount of information about their students. Teachers usually seek information about their students' general abilities or achievement, class participation, self-concepts, social competence, independence, classroom behavior, and work habits (Shavelson & Stern, 1981). This information comes from many sources, such as their own informal observations, anecdotal reports of other teachers, standardized test scores, and school records. In order to use a large amount of information, teachers integrate it to form judgments about students' cognitive, affective, and behavioral states. (p. 397)

Teachers are considered to make judgments and carry out decisions in an uncertain and complex environment (Shavelson & Stern, 1981; Shulman & Elstein, 1975). In order to handle this complexity, teachers should develop some strategies or procedure in the face of miscellaneous information of students.

The lens model developed by Brunswik (1955) has helped to understand and externalize the judgment process. The lens model is generally composed of three stages: the

true state of target, the perceivable attributes of the target, and the judged state of the target. To predict specific students' aspects, teachers identify information they think to be related and incorporate it into judgments. For example, for grading students in mathematics, teacher candidates based their judgments on information regarding students' German achievement as well as their general intelligence (Kaiser, Möller, Helm, & Kunter, 2015).

1.3.2 Student Characteristics Influencing Teacher Judgment

Teachers' judgments of students are based on various sources of information. This information includes diverse student characteristics (Bressoux & Pansu, 2016). Student demographic characteristics, e.g., facial attractiveness, parents' education, student gender (Baudson, Fischbach, & Preckel, 2016; Dusek & Joseph, 1983; Holder & Kessels, 2017), scholastic characteristics, e.g., the academic development, the grade point average, the quality of students' work (Doherty & Conolly, 1985; Praetorius, Koch, Scheunpflug, Zeinz, & Dresel, 2017; Rich, 1975), and behavior in class, e.g., interaction with teachers, students' bad behavior, teacher-student relationships (Bennett, Gottesman, Rock, & Cerullo, 1993; Hecht & Greenfield, 2002; Rubie-Davies, 2010; Timmermans, de Boer, & van der Werf, 2016), have been identified as influencing the accuracy of teachers' judgments.

Although there has been some research showing that teachers' judgments are based on relevant student information, the influence of many other student characteristics is still rarely investigated and remains largely inconclusive (Baudson et al., 2016; Jussim & Harber, 2005; Meissel, Meyer, Yao, & Rubie-Davies, 2017; Oudman, van de Pol, Bakker, Moerbeek, & van Gog, 2018; Südkamp et al., 2012). A holistic examination of students' characteristics will further help to clarify which information leads to an accurate judgment and which information leads to stereotypes. The lens model could help to gain more insight into teacher judgment processes.

Meanwhile, methodological and perspective differences make the results not comparable across studies. There are considerable differences in studies of correlates of

teacher judgment. Methodological differences arise from the use of quantitative or qualitative research approaches, whereas perspective differences result from the theoretical point of view. The differences in research approaches should be briefly outlined.

1.3.3 Methodological and Perspective Differences in Studies of Correlates of Teacher Judgment

Existing research has frequently used quantitative approaches to examine the association of some specific factors and teacher judgments. For example, Kaiser, Retelsdorf, Südkamp, and Möller (2013) ran a structural equation model to document an effect of student engagement on teachers' judgments of student achievement. In two longitudinal studies, Chamorro-Premuzic and Furnham (2003) showed that student personality is significantly related to academic performance. However, it was also figured out that teachers do not account for individual differences in students' personality when predicting their final grades. In addition, Hecht and Greenfield (2002) determined by quantitative analysis the role of gender, classroom behavior, and emergent literacy skills in teacher judgment.

Some studies have applied qualitative methods to explain teachers' reasoning behind their judgments. He, Valcke and Aelterman (2012) asked in-service teachers to define their evaluation beliefs in a semi-structured interview. Wijnia, Loyens, Derous, and Schmidt (2016) used a qualitative approach and found that university teachers built their judgments upon the observations of university students' engagement and motivation. Besides, St-Onge, Chamberland, Lévesque, and Varpio (2016) qualitatively investigated raters' cognitive process while assessing examinee's clinical performance displayed in a video. They found that raters relied on both external (such as examinee's performance or outside standards of performance) and internal sources of information (such as their own standards of performance for a given trainee level).

In addition, research focusing on teacher judgment accuracy and its variability has tended to explain the influencing factors from different theoretical perspectives. For example,

some researchers have examined bias or stereotypes in teachers' judgments. The study by Holder and Kessels (2017) argued that German student teachers showed gender and ethnic bias when estimating a fictitious student's actual performance on an objective scale. Kaiser et al. (2015) found that teacher candidates graded students in mathematics based on information regarding student intelligence and German achievement. No bias was found towards family background and self-concept.

Another perspective on achievement judgment is information utilization. Kishor (1994) explored how teachers mentally use performance information in judging their students. Based on Kelley's (1967) model of causal judgment, his study categorized students' performance data into consensus, consistency, and distinctiveness information. Analyses revealed that teachers mainly relied on consensus information for making diagnostic and predictive judgments.

Altogether, the diverse literature suggests that various student characteristics influence teachers' judgment process and both quantitative and qualitative approaches have been applied to study the associated factors. Nevertheless, the existing studies about teacher judgment processes look relatively scattered and immethodical. Studies that summarize and structure all related student information and examine teachers' judgment strategies from their perception of student information are expected to come.

1.4 The Improvement of Teacher Judgment Accuracy

Teachers' judgments are of enormous significance; however, their judgment accuracy was shown to be far from perfect (Hoge & Coladarci, 1989; Südkamp et al., 2012). Another important perspective for forthcoming studies is to examine the possibilities of improving teacher judgment accuracy.

1.4.1 Promoting Teachers in the Judgment Process

According to the lens model of Brunswik (1955), teachers could make reliable predictions of students' achievement when they have access to information with a high

correlation to students' actual performance. Findings of previous research also indicate that teachers' knowledge on students' overall performance plays an essential role in their judgment process (Glogger-Frey et al., 2018; Oudman et al., 2018). Therefore, judgment accuracy will improve when teachers are provided with more useful student information.

Although there are a variety of student cues, teachers' judgments were reported to rely more on available, memory-based than continuously updated information (Hattie & Timperley, 2007; Shavelson, 1983). The lack of frequent and objective information about students' current status makes it a challenge for teachers to give precise judgments.

1.4.2 Empirical Evidence for Improving Teacher Judgment

Given the different moderators to determine teacher judgment accuracy, there has been some empirical evidence shown that accuracy of teacher judgment is malleable (Klug, Gerich, & Schmitz, 2016; Thiede et al., 2015, 2018; Trittel, Gerich, & Schmitz, 2014).

Among all the intervention studies, teacher professional training program is one of the most important approaches for improving teacher judgment competence (Klug, Gerich et al., 2016; Thiede et al., 2015, 2018). Judgment accuracy was greater for teachers who participated in the *Developing Mathematical Thinking* (DMT) professional development focused on improving student-centered mathematics instruction (Thiede et al., 2015, 2018). Moreover, both a training program (Klug, Gerich et al., 2016) and a hands-on seminar (Trittel et al., 2014) on educational diagnostics for prospective teachers provided opportunities to promote teachers' diagnostic competence.

In addition to gaining knowledge of judgment methods and making more classroom practices, another likely explanation for the improved judgment accuracy in these training programs is that teachers were informed with valid student information (Thiede et al., 2015, 2018; Trittel et al., 2014). For the effect of DMT, it was hypothesized that student-centered teaching would promote teacher-student conversations, and this increases teachers' awareness of cues that are diagnostic of student learning. In the end, judgment accuracy for teachers in

the DMT professional development group was greater than for teachers in other groups. In another study from Trittel et al. (2014), the availability of instant and objective student information might also have contributed to the fostered diagnostic competence. Instead of training programs, which rely on interpersonal interactions, a new approach of using technology in the classroom was taken into consideration.

1.4.3 The Use of Learner Response Systems in the Classroom

Learner response systems, or "clickers", can be defined as instructional technologies that allow teachers to rapidly collect and analyze student responses to questions posed during class (Bruff, 2009). Clickers are increasingly used to track students' study in the classroom and the interaction between teachers and students. From empirical studies over the last two decades, the use of clickers has gained widespread acceptance and recognition, leading to positive student learning outcomes in the classroom (Anderson, Healy, Kole, & Bourne, 2013; Hunsu, Adesope, & Bayly, 2016; Keough, 2012; Mayer et al., 2009). In particular, information gathered by teachers and immediate feedback to students were frequently utilized to explain how learner response systems contribute to the improvement of teaching and learning processes (Chien, Chang, & Chang, 2016; Faber, Luyten, & Visscher, 2017; Lantz & Stawiski, 2014). It is suggested that teachers could also incorporate the direct feedback about students' learning outcomes into their judgment process. Hence, the use of clickers in the classroom would shed light on opportunities for improvement of teacher judgment accuracy.

1.5 The Stability of Teacher Judgment Accuracy

1.5.1 The Repeatability of Existing Studies

The variability of teacher judgment accuracy facilitated studies to explore possible explanations (Paleczek et al., 2017; Südkamp et al., 2012). It has been shown that the variability of judgment accuracy is in connection with various judgment, test, teacher, and student characteristics. Teacher judgment accuracy will be influenced when any of the four

conditions change. However, it is also interesting to see the repeatability of teacher judgment accuracy in the same measurement setting over a period of time.

The results of each study could only reflect teachers' judgment accuracy at a certain time, considering that most of the existing studies were designed cross-sectionally instead of applying a repeated measurement design. Consequently, it is difficult to figure out the judgment accuracy of each individual teacher over time. Specifically, whether a teacher could judge with the same accuracy in a different situation or at different times is still unclear.

1.5.2 Teacher Judgment Accuracy over Time

So far, there are only a few longitudinal studies that have looked at the changes of teacher judgment accuracy. For example, Lorenz and Artelt (2009) examined the diagnostic skills of elementary school teachers within a time interval of six months. In the areas of vocabulary, text comprehension and arithmetics, teachers were able at both times to predict student achievement on standardized tests with moderate accuracy. Rank component differences between measurement points were not significant. Another longitudinal study by Hinnant, O'Brien and Ghazarian (2009) covered a period of four years. Teachers were asked to rate reading and mathematics abilities in the first, third, and fifth grade. Moderate correlations in both subjects were found at all times, which did not deviate significantly from each other. The disadvantage with these studies is that self-fulfilling prophecy effects are hard to be ruled out. Moreover, there is almost no research involved to measure the changes of teacher judgment accuracy on student motivation and emotions. The temporal stability of teachers' judgment accuracy on both students' cognitive and affective aspects would be the last question to be answered.

1.6 Aims of the Studies

Based on the review of previous studies, a systematic analytical framework which consists of three main studies was prepared to broaden the understanding of teacher judgment. The focus of Study 1 was to examine teachers' judgments of student achievement and their

strategies behind the judgments. In this study, the lens model of Social Judgment Theory provided a framework for understanding how teachers form their judgments and why judgment accuracy varies between individuals. Specifically, different student cues that are available to teachers were identified and categorized. These were incorporated into a semi-structured questionnaire to investigate how teachers select and use information from a large variety of student data.

Study 2 was devoted to promoting teachers in the judgment process. In addition to the research on teacher judgment accuracy and its influencing factors, attempts to promote teachers effectively are still in the early stages. This study aimed to enhance judgment accuracy by providing teachers with more information about students' learning outcomes by the use of learner response systems (clickers) in the classroom. The improvement of teacher judgment accuracy about student achievement was measured in a pre-post-test intervention study to examine the effects of clickers in the classroom. Moreover, it was checked whether the regular use of clickers resulted in higher student achievement.

Finally, the temporal stability of teacher judgment accuracy about student achievement, motivation, and emotions was further explored in Study 3. Using a time interval of four weeks, this study was able to investigate to what extent teacher judgment accuracy remains temporally stable for both cognitive and motivational-affective student characteristics. In addition, it examined the interplay of various motivational, emotional, and cognitive factors to holistically explore the accuracy of teacher judgments of student characteristics.

CHAPTER 2: EXAMINING TEACHERS' STRATEGIES TO JUDGE STUDENT ACHIEVEMENT FROM A CUE UTILIZATION PERSPECTIVE

2.1 Introduction

The accuracy of teacher judgment is the basis for a fair evaluation of student achievement. Only if teachers correctly include and combine all the necessary information in the judgment process, students can hope for a fair assessment of their academic achievement. Empirical research of the past decades has shown that teachers are able to make relatively accurate judgments of student achievement; even though, these judgments are far from perfection (Hoge & Colardarci, 1989, Südkamp, Kaiser, & Möller, 2012).

While teacher judgment is fairly accurate but by no means exact, it is important to understand teachers' strategies behind the judgments, especially what information they are using in the judgment process. Educational research therefore often tested whether teacher's judgment on student achievement is systematically influenced by other factors (Holder & Kessels, 2017; Kaiser, Möller, Helm, & Kunter, 2015). These may include demographic information, student abilities and attitudes, in and out of class behavior, past academic performance, homework, or social interactions with parents and other teachers. All of these sources can provide important information about student achievement that may be incorporated into teachers' judgment strategies.

The purpose of this study is to learn more about the strategies that teachers are using to gauge student achievement. Hattie (2012), with his compilation of meta-analyses, has shown that a wide range of factors is influencing students' academic achievement. What factors, however, do teachers make of use of to arrive at accurate judgments of student achievement? In this study, it should be examined if and to what extent central indicators of student achievement (Hattie, 2012) are incorporated into teachers' judgment process. In other words, information about the richness of distinctive factors should be obtained, which help teachers to make the best possible judgments about student achievement.

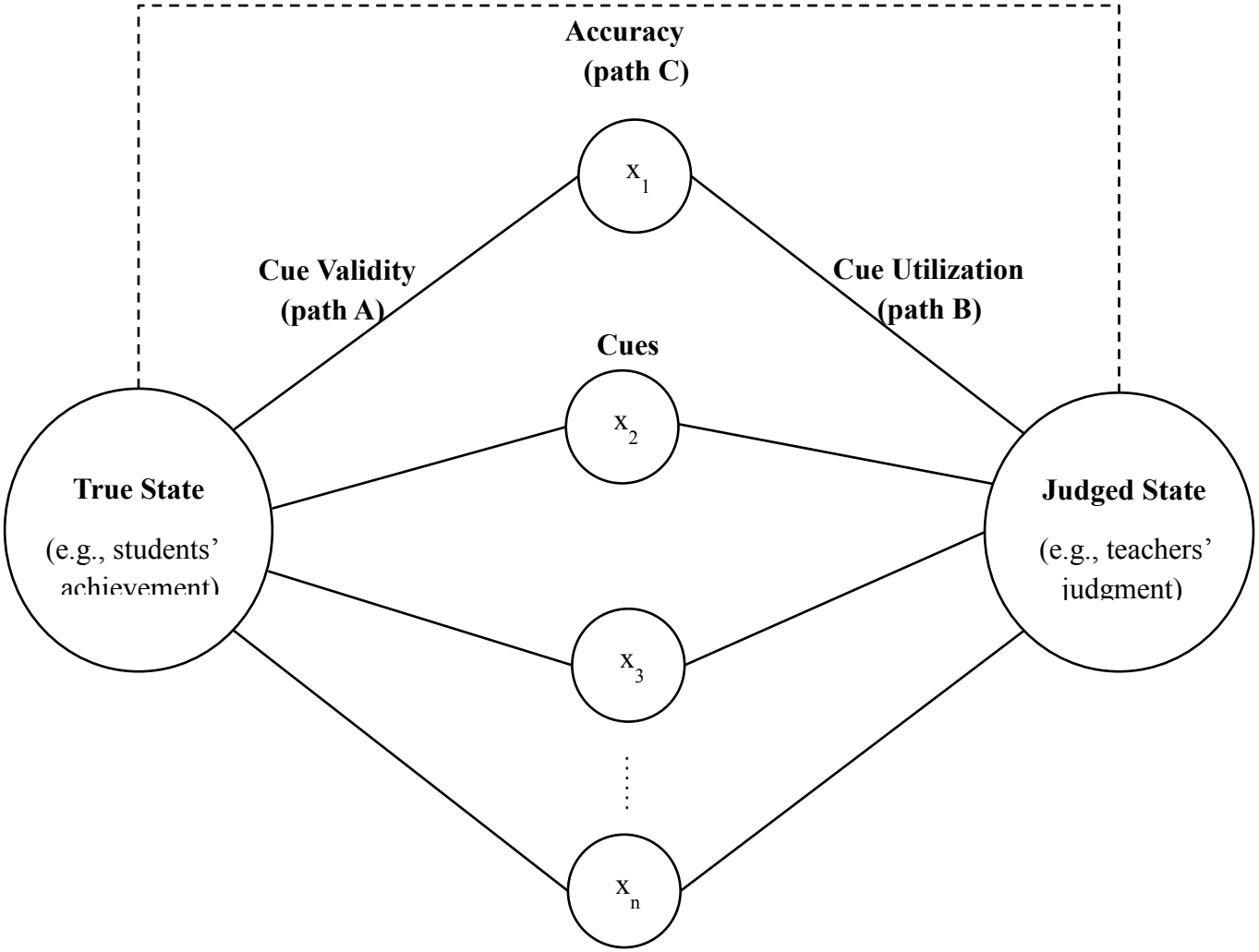
2.1.1 Understanding Teachers' Judgment through Social Judgment Theory

Social judgment theory (SJT) can be regarded as a theoretical framework for describing and understanding the formation of teachers' judgment. It was initially derived from the study of human judgment in social situations (Hammond, Rohrbaugh, Mumpower, & Adelman, 1977). In the educational context, researchers have used SJT to model how judgments are made by teachers in educational settings in order to understand and improve important judgment processes (Cooksey, Freebody, & Davidson, 1986; Haigh, Ell, & Mackisack, 2013; Haigh & Ell, 2014).

The essential paradigm of SJT can be embodied in the lens model (Brunswik, 1955; Cooksey et al., 1986) (see Figure 2-1). The lens model is generally composed of three parts: (1) the true state of the target is presented on the left side, i.e., in the current study, students' real achievement. (2) Various cues are presented in the center of the lens diagram. These cues are a set of perceivable attributes available to teachers to form a judgment, e.g., students' demographic information. (3) The judged state of the target is presented on the right side of the model, i.e., teachers' judgments of student achievement. The extent to which a cue is in fact related to the actual state is called cue validity (path A). The strength with which a teacher does in fact incorporate a cue into judgments is called cue utilization (path B) (Nestler & Back, 2013). Cue validity and cue utilization conjointly reflect whether teachers' judgments are associated with valid or misleading information about the true state (e.g., students' achievement). The more teachers rely on information (cue) with a high correlation to students' actual achievement, the more likely they can make reliable predictions (path C). Therefore, the cues in the analysis framework are playing a crucial role in explaining judgment accuracy. It suggests that investigations on how teachers select and use the information from a large variety of student data are of great significance for identifying teachers' judgment strategies. Yet, studies which associated teachers' judgment process with the use of student information are quite miscellaneous, relatively fragmented, and most of

them were implemented in a quantitative way with relevant data not consistently reported (Oudman, van de Pol, Bakker, Moerbeek, & van Gog, 2018; Südkamp et al., 2012).

Figure 2-1. Model of Social Judgment Theory to Explain the Accuracy of Judgment between True and Judged State



2.1.2 Validity of Different Student Information

Cues with high validity for judging students' achievement are supposed to be highly correlated with students' real performance. Actually, there is a vast number of studies that have examined correlates of students' academic performance. To better understand the formation of teacher judgment in the lens model, a parsimonious, yet comprehensive overview of the correlates of students' academic achievement (path A in Figure 2-1) is presented in Table 2-1. A wide range of information sources on the level of students was categorized into the following domains: (a) students' abilities and attitudes, (b) behavior during class, (c) tests, (d) homework, (e) behavior after class, (f) demographics, and (g) other social interactions. The order of findings in Table 2-1 resembles the order of categories of teachers' judgment strategies (path B in Figure 2-1) under investigation.

Table 2-1

Meta-analytic Relations between Student Characteristics and Academic Achievement

Authors	Year	*k*	*r*	Predictor variable	Criterion variable
			Abilities and attitudes		
Roth, Becker, Romeyke, Schäfer, Domnick, & Spinath	2015	240	0.54	Intelligence	School grades
Schiefele, Krapp, & Schreyer	1993	21	0.30	Interest	Academic achievement
Dent & Koenka	2016	61	0.11	Cognitive strategies	Academic achievement
Cerasoli, Nicklin, & Ford	2014	183	0.26	Intrinsic motivation	Performance
Hulleman, Schrager, Bodmann, & Harackiewicz	2010	243	-0.13 to 0.11	Four achievement goals	Performance outcomes
Huang	2012	151	-0.13 to 0.13	Four achievement goals	Academic achievement
Talsma, Schütz, Schwarzer, & Norris	2018	11	0.25	Self-efficacy	Academic performance
Huang	2011	39	0.24 to 0.25	General self-concept	Subsequent academic achievement
Möller, Pohlmann, Köller, & Marsh	2009	69	0.49 to 0.61	Domain-specific self-concept	Domain-specific achievement
Petscher	2010	32	0.32	Attitude to reading	Achievement in reading
Ma	1999	26	-0.27	Anxiety towards mathematics	Achievement in mathematics
			Behavior during class		
Lei, Cui, & Zhou	2018	69	0.27	Engagement	Academic achievement
Gray, Dueck, Rogers, & Tannock	2017	27	-0.15 to -0.64	Teacher-rated inattention	Academic achievement
Lei & Cui	2016	35	0.31	Positive high arousal (Enjoyment, hope, pride)	Academic achievement
			-0.37	Negative low arousal (hopelessness, boredom, depression, exhaustion)	Academic achievement
Tze, Daniels, & Klassen	2016	29	-0.24	Boredom	Academic outcomes
			Tests		
Duncan et al.	2007	6	0.34	School-entry mathematics	Later achievement
			0.17	School-entry reading	Later achievement
Richardson, Abraham, & Bond	2012	1105	0.01	Pessimistic attributional style	Academic performance

Singer & Strasser	2017	68	0.55	Arithmetic performance	Reading performance
von der Embse, Jester, Roy, & Post	2018	238	-0.16 to -0.27	Test anxiety	Achievement tests
			Homework		
Fan, Xu, Cai, He, & Fan	2017	61	0.22	Mathematics homework	Performance in mathematics
Cooper, Robinson, & Patall	2006	32	0.18	Time spend on homework	Mathematics achievement
			0.12	Time spend on homework	Reading achievement
Kim & Seo	2015	33	-0.13	Procrastination	Academic performance
			Behavior after class		
Roorda, Koomen, Spilt, & Oort	2011	99	0.16	Positive teacher-student relationship	Achievement
Lauer, Akiba, Wilkerson, Apthorp, Snow, & Martin-Glenn	2006	30	0.07	Out-of-school programs on reading	Reading achievement
			0.09	Out-of-school programs on mathematics	Mathematics achievement
Murayama & Elliot	2012	81	0.02	Structural competition	Performance
Dietrichson, Bøg, Filges, & Klint Jørgensen	2017	101	0.18	Tutoring	Educational Achievement
			Demographics		
White	1982	101	0.31	Socio-economic status	Achievement
Sirin	2005	58	0.29	Socioeconomic status	Academic achievement
Voyer & Voyer	2014	369	-0.11	Gender (Boys vs. girls)	Scholastic Achievement
Lindberg, Hyde, Peterson, & Linn	2010	242	0.03	Gender (Boys vs. girls)	Mathematics Performance
Lietz	2006	139	-0.10	Gender (Boys vs. girls)	Reading achievement
Malouff & Thorsteinsson	2016	20	0.30	Physical attractiveness	Subjective grading
			Social interaction		
Castro, Expósito-Casas, López-Martín, Lizasoain, Navarro-Asencio, & Gaviria	2015	37	0.12	Parental involvement	Academic achievement
Fan & Chen	2001	25	0.25	Parental involvement	Academic achievement

Note. *k* = number of studies / independent samples.

Table 2-1 presents the results from 34 meta-analyses with details about the number of studies resp. independent samples (*k*) and effect sizes in terms of Pearson correlations (*r*). The meta-analyses mainly focus on the area of K-12 education with respect to the subjects of language arts and mathematics. Due to a lack of data in some meta-analyses, details of the sample size were not reported here. In order to enable a comparison between meta-analyses, effect sizes of Cohen's *d* statistic or Hedges' *g* were transformed into *r*. According to Cohen's (1992) guidelines, an effect size of $r = .10$ means small, $r = .30$ means medium, and $r = .50$ means large.

Some student cues are quite valid indicators of student achievement. Intelligence ($r = .54$) and domain-specific self-concept ($r = .49$ to $.61$) were the strongest predictors of student achievement in the domain of abilities and attitudes. Teacher-rated inattention showed in some cases strong negative correlations ($r = -.15$ to $-.64$) with academic achievement. Students' test information, especially arithmetic performance revealed a large effect ($r = .55$) on reading performance. On the other side, student information like pessimistic attribution style ($r = .01$) and structural competition ($r = .02$) do not seem to be consistently related to academic achievement. Also other effects of behavior after class, for example, out-of-school-programs on reading ($r = .07$) and mathematics performance ($r = .09$), show relatively weak impact on student achievement.

2.1.3 Student Information Sources of Teacher Judgment

According to the lens model of Brunswik (1955), the more valid information teachers are using, the more accurate judgments they can make. Therefore, it is necessary to investigate whether and how teachers attach significance to diverse student information in order to form their judgments.

Abilities and Attitudes

Teachers make use of different information from students' abilities and attitudes to judge their achievement. A meta-analysis by Hoge and Coladarci (1989) showed that teachers'

judgments of students' achievement were influenced by students' academic ability, for example, teachers displayed a tendency to overestimate the performance of highly intelligent students. Moreover, students' motivational-affective characteristics like learning motivation and self-confidence were found to have an influence on teachers' expectations for students' learning outcomes (Rubie-Davies, 2010; Timmermans, de Boer, & van der Werf, 2016; Urhahne, 2015).

A meta-analysis from 33 studies (Machts, Kaiser, Schmidt, & Möller, 2016) reported a mean judgment accuracy of cognitive abilities of $r = .43$, and an intelligence judgment accuracy of $r = .50$. However, some other studies also revealed that teachers do not know much about students' motivational-affective traits such as learning motivation and test anxiety (Karing, Dörfler, & Artelt, 2015; Urhahne et al., 2011), while both students' cognitive and motivational-affective characteristics are important predictors of further learning (Jurik, Gröschner, & Seidel, 2013; Marjoribanks, 1987; Snow, Corno, & Jackson, 1996).

Classroom Behavior

Students' classroom behavior, especially the interaction with teachers, is reflected in teacher ratings of students' academic performance (Bennett, Gottesman, Rock, & Cerullo, 1993; Hecht & Greenfield, 2002; Timmermans et al., 2016). According to the study of Bennett et al. (1993), teachers' perceptions of students' behavior constituted a significant component of their academic judgments. It was found that teachers assume lower academic performance when they detect students exhibiting bad behavior. In the same vein, Rubie-Davies (2010) demonstrated that teachers' expectations of academic performance are closely related to the perceived teacher-student relationships. However, other findings show that ratings of reading achievement based on students' classroom behavior led to lower accuracy (Hecht & Greenfield, 2002). Thus, it seems worthwhile to examine how teachers think about students' classroom behavior and how their perceptions of classroom behavior are reflected in their achievement judgments.

Tests

Students' test results and records have usually been taken as a reliable predictor for students' future academic performance. For instance, Rich (1975) reported that elementary school teachers' judgments partly resulted from the provided information about the academic development of a child. Similarly, the GPA which represents students' grade point average across different subjects was found to be related to higher levels of judgment (Praetorius, Koch, Scheunpflug, Zeinz, & Dresel, 2017). Furthermore, evidence from a meta-analysis by Hoge and Coladarci (1989) suggested that it might be easier for teachers to accurately assess high-performing than low-performing students.

However, in addition to test results, students' test-relevant motivation and emotions (e.g., test anxiety) are expected to be related to learning outcomes and teachers' diagnostic skills (Westphal, Kretschmann, Gronostaj, & Vock, 2018). Thus, although prior test performance is assumed to be largely involved in achievement-related judgments, test-relevant motivation and engagement of the students can be potential information related to teachers' concerns.

Homework

The vast majority of studies about homework has focused on the homework-achievement relationship (Cooper, Robinson, & Patall, 2006; Fan, Xu, Cai, He, & Fan, 2017). For example, a highly influential meta-analysis from Cooper et al. (2006) reported a positive effect of homework on achievement. A newer meta-analysis of Fan et al. (2017) found a similar result with an overall small and positive relationship between homework and academic achievement in mathematics and science.

In contrast to the important role that homework is playing in students' academic achievement, very little research has taken homework into consideration when trying to explore the factors that may affect teachers' evaluation. To our knowledge, only a single study by Doherty and Conolly (1985) has examined the relation between the quality of

children's written work and teachers' estimated scores in standardized tests of attainment. Findings showed a very significant correlation between teacher scores and their impression on the tidiness of written work. Furthermore, teacher scores were significantly correlated to students' actual achievement, which means teachers can judge quite accurately when they rely in their judgments on the quality of students' work.

Demographics

Student demographic information is one of the most often used information in prior studies to examine the influence on teachers' judgments. Extensive research revealed effects of some student characteristics on teachers' judgment accuracy. For example, students with facial attractiveness were found to have an influence on teachers' assumption of academic performance (Dusek & Joseph, 1983). Baudson, Fischbach, and Preckel (2016) demonstrated that parents' education level strongly affected teachers' judgments of students' cognitive ability. Empirical findings on student gender, however, were inconsistent. Some researchers found that teachers' judgments are biased by gender stereotypes (Baudson et al., 2016; Holder & Kessels, 2017). For example, it was concluded that teachers tend to rate boys lower than girls in both mathematics and reading achievement. However, other studies could not corroborate such differences (Bennett et al., 1993; Paleczek, Seifert, & Gasteiger-Klicpera, 2017).

Moreover, some demographic information about students is confirmed to be related to their achievement. A meta-analysis from Richardson, Abraham, and Bond (2012) showed that students from higher socio-economic background, students of older age, and female students obtain higher grades. While another meta-analysis by Reilly, Neumann, and Andrews (2015) demonstrated that boys, in general, outperform girls in mathematics.

Social Interactions

Teacher-parent communication as a kind of social interaction of teachers was found to have a positive influence on student achievement (Kraft & Rogers, 2015; Pang & Watkins,

2000). Pang and Watkins (2000) explained that teachers and parents were allowed to exchange information and ideas about the learning situation of students. This provides teachers with a comprehensive understanding of students' attitudes towards learning and their difficulties, and enables teachers to better support students at school. Thus, although social interactions provide teachers with helpful student information, little is known about whether teachers' judgments of students are influenced by their interactions with parents and other colleagues.

Taken together, prior studies have shown that teachers probably have acquired and used information from different sources to make judgments about students' learning outcomes. The knowledge about the use of these cues, however, is fragmented and incomprehensive. Systematic investigations of the utilization and validity of different types of cues in a more qualitative manner is lacking.

2.1.4 The Current Study

Study Context

Chinese teachers possess a wide spectrum of student information sources as they are—in addition to daily classroom teaching—responsible for many student-related activities (Chen, 2019). Due to the high pressure for academic performance, teachers are considered to pay great attention to students' homework and tests (Kim & Fong, 2013). Therefore, after class teachers often stay in the classroom and correct homework or tests in order to offer students direct feedback about their mistakes. Nowadays, the use of social media also enables teachers to talk with parents about students' homework and learning outcomes at school. It is also likely that teachers advise students to attend tutoring after school to support their learning. To promote students' interest and achievement, Chinese students are encouraged to take part in discipline competitions (e.g., Mathematics Olympics), and teachers are usually involved as their instructors. Moreover, teachers are even responsible for some administrative tasks such

as distributing lunch. It helps them to build a close relationship with students and know more than teachers in other countries.

Given the specific cultural values of emphasising teachers' academic accountability and students' achievement, studies that summarize and structure all related student information and examine teachers' judgment strategies from their perception of student information are expected to come.

Study Aims

The main goal of this study is to specify what student information teachers are using to judge student achievement. The study itself can be divided into two parts which will be discussed separately for the sake of clarity.

The first part is dealing with the problem to identify and define different types of student information that are available to teachers. In this part, 34 meta-analyses and a previous interview with 16 primary school teachers have been reanalyzed. It turned out that mainly seven student information sources are obtainable to teachers. These sources offering information about the validity of student cues were presented in the introduction.

The second part encompasses the body of research. Through the use of semi-structured questionnaires, it was investigated if and to what extent teachers utilize these seven types of student information sources. The specific research questions for this study were as follows:

1. What cues from the seven student information sources, do teachers select as relevant in order to judge student achievement?

2. What cues from the seven student information sources, do teachers rank as most important to judge student achievement?

3. Which of the seven student information sources do teachers rank as most important to judge student achievement?

2.2 Method

2.2.1 Participants

The sample consisted of N = 260 teachers (87.3% female) from seven Chinese primary schools. Schools were selected from three eastern coastal cities with high education quality and economic growth. All teachers took part on a voluntary basis and did not receive any further gratification. Teachers' mean age was 36.83 years (SD = 8.13) and they had an average teaching experience of 15.45 years (SD = 9.49). No teacher had been teaching for less than one year. Teachers were responsible for educating students in Chinese (n = 82, 31.5%), English (n = 65, 25.0%), or mathematics (n = 113, 43.5%) as their only subject in primary school. About half of them were working as class teachers (n = 116, 44.6%).

2.2.2 Materials

A semi-structured questionnaire was developed for the purpose of this study. The structure and items of the questionnaire were obtained from the findings of the meta-analyses (see Table 2-1) in combine with data from qualitative interviews of Chinese primary school teachers (Zhu, 2014).

In one-to-one interviews, 16 English teachers at Chinese primary schools were asked to give reasons for their judgments on students' English achievement. They provided different information about the sources relevant for their judgment process. Relying on the interview records, teachers' statements reflecting their channels for collecting information were progressively coded into the seven broad categories. At the same time, taken the much-noticed work of Hattie (2012) as a basis, meta-analyses were identified that offer information on the validity of these student information cues.

Through this combined top-down and bottom-up approach, the final semi-structured questionnaire with seven categories and accompanying items was developed. The semi-open questions provided teachers with choices to indicate all of their information sources to properly judge student achievement. For example, the first category "student abilities and attitudes" consisted of eight items that were often mentioned in the teacher interviews and could have been substantiated through the meta-analyses. Each semi-open question contained

a further item "other" in case a judgment option was missing. Some utterances of the teachers could not be exactly covered by the meta-analyses, but were so frequently raised in the interviews that the decision was made to include them as items of the questionnaire. The wording of the items was intended to express the constructs under investigation in close relation to teachers' work context.

The final semi-structured questionnaire consisted of ten semi-open questions with diverging number of items which built the core of the questionnaire (e.g., what information about student behavior during class helps you to judge student achievement? Please mark the factors that you use and rank the top three). Seven closed questions to gather teachers' demographic information and two open-ended questions (e.g., how do you ensure that judgments about student achievement are soundly based?) completed the questionnaire.

2.2.3 Procedure

The semi-structured questionnaire was carried out with teachers after school during an investigation period of two weeks. It was conducted by six trained Chinese-speaking student helpers. Teachers in each school were informed by the principal about the investigation and were invited to fill in the questionnaire independently lasting about 25 minutes. Teachers were asked to point out what information in each category they rely on and make an order of the most important three items. Teachers also had the option of not selecting any of the listed information and assigning a rank to it if the items did not met their ideas.

2.2.4 Statistical Analyses

Multivariate analyses and non-parametric statistical procedures were used to analyze the semi-open questions. To examine teachers' selection of different information sources, a multiple response analysis was applied for the frequency of choices. One-way repeated measures ANOVA and subsequent pairwise comparisons with Bonferroni correction enabled to test for significant differences in teachers' choices. For teachers' ranking, the rank response analysis presented frequencies of each sub-category from the most important to the third

important. In addition, an average rank was built for further analysis. In order to determine the average rank, non-ranked items of each teacher were replaced by expectancy values which are determined by the mean rank of these items. Friedman's ANOVA's and subsequent Wilcoxon signed rank tests with Bonferroni correction were carried out to test differences between teachers' ranking of information.

Descriptive statistics generated from the closed questions were used for describing teachers' basic characteristics. Teachers' statements in the open-ended questions about their metacognitive judgment processes were recoded as text and analyzed in a recursive process (Bos & Tarnai, 1999). They will not be presented as a main finding but as a supplement to the interpretation of the results.

2.3 Results

2.3.1 Selection and Ranking of Student Abilities and Attitudes

Table 2-2 displays teachers' selection and ranking of students' abilities and attitudes. The selection equals the frequency with which the eight types of student data were mentioned by the teachers when they judge student achievement. Results of a one-way repeated measures ANOVA (Wilks-Lambda = .173, $F(7, 252) = 171.59$, $p < .001$, $\eta_p^2 = .827$) indicate significant differences among the selected eight information types. All types of data were used. However, subsequent pairwise comparisons with Bonferroni correction showed that general intelligence, interest and learning strategies were used most often, whereas students' verbal skills and anxiety about the subject were hardly mentioned.

In a similar way, teachers' rank order of information is in line with their selection. Friedman's ANOVA shows significant differences between the ranks of the eight items ($\chi^2(7) = 453.55$, $p < .001$). Again, subsequent Wilcoxon signed rank tests with Bonferroni correction reveal that general intelligence, interest and learning strategies were given the highest priority, while verbal skills and anxiety about the subject were considered least important.

It is notable that subject differences were found in the selection of mathematical and verbal skills. Language teachers utilized verbal skills more often, while mathematics teachers chose mathematical skills with higher frequency ($p < .001$).

Table 2-2

Teachers' Selection and Ranking of Abilities and Attitudes to Judge Student Achievement

	Selection		Ranking				
	n	%	*n*	First	Second	Third	Average rank[†]
General intelligence	200	77.2[a]	176	79	44	53	3.19[a]
Interest	191	73.7[a]	161	59	56	46	3.47[ab]
Learning strategies	186	71.8[ab]	158	27	63	68	3.73[ab]
Motivation	153	59.1[b]	118	57	30	31	4.08[b]
Self-confidence	106	40.9[c]	57	10	27	20	5.16[c]
Mathematical skills	78	30.1[c]	45	9	18	18	5.34[c]
Verbal skills	49	18.9[d]	18	1	8	9	5.75[d]
Anxiety about the subject	28	10.8[d]	5	0	2	3	5.93[d]

Note. Different superscripts indicate significant differences between categories. [†] Each teacher could rank up to 3 items. In order to determine the average rank, non-ranked items of each teacher were replaced by expectancy values.

2.3.2 Selection and Ranking of Behavior during Class

Teachers' selection and ranking of six types of information about students' behavior during class are displayed in Table 2-3. One-way repeated measures ANOVA shows significant differences between items (Wilks-Lambda = .46, $F(5, 254) = 59.35$, $p < .001$, $\eta_p^2 = .539$). The majority of statements concern students' concentration during class. More than half of the considerations are related to students' other behavior, including raising hands, joining classroom activities, having passion for the class, and communicating with teachers.

Rank differences between various types of information were getting significant ($\chi^2(5) = 344.018$, $p < .001$). Teachers' perceptions of the importance of information were quite consistent with their selection.

Table 2-3

Teachers' Selection and Ranking of Behavior during Class to Judge Student Achievement

	Selection		Ranking				
	n	%	*n*	First	Second	Third	Average rank[†]
Concentrates well	242	93.4[a]	226	161	44	21	1.78[a]
Raises hands often	175	67.6[b]	126	11	52	63	3.67[b]
Likes to join classroom activities	164	63.3[b]	107	20	47	40	3.77[b]
Has passion for the class	158	61.0[b]	125	28	55	42	3.53[b]
Communicates well with me	144	55.6[b]	94	8	36	50	3.99[bc]
Is prepared for class	106	40.9[c]	60	18	14	28	4.27[c]

Note. Different superscripts indicate significant differences between categories. [†] Each teacher could rank up to 3 items. In order to determine the average rank, non-ranked items of each teacher were replaced by expectancy values.

2.3.3 Selection and Ranking of Test Information

With respect to tests, results of a one-way repeated measures ANOVA indicate statistically significant differences among the six types of information (Wilks-Lambda = .212, $F(7, 253) = 188.213$, $p < .001$, $\eta_p^2 = .788$). Students' last test performance was selected as the most often used information when teachers make predictions about student achievement. Moreover, three types of information, including attribution of failure, test strategies, and academic files, show a high percentage of use. Test anxiety was much less influential in teachers' judgments.

A Friedman's ANOVA test shows significant rank differences of the test information ($\chi^2(5) = 466.531$, $p < .001$). Teachers considered the last test performance that they mainly relied on as the most important, whereas the least marked information, test anxiety, has the lowest ranking. These findings again suggest that teachers believed they are supported by effective information to make sound and accurate judgments (see Table 2-4).

Table 2-4

Teachers' Selection and Ranking of Tests to Judge Student Achievement

	Selection		Ranking				
	n	%	*n*	First	Second	Third	Average rank[†]
Last test performance	240	93.0[a]	225	166	40	19	1.76[a]
Attribution of failure	162	62.8[b]	139	21	41	77	3.50[b]
Test strategies	161	62.4[b]	144	34	83	27	3.21[b]
Academic file	154	59.7[b]	129	23	53	53	3.53[b]
Grades of other subjects	84	32.6[c]	63	1	19	43	4.33[c]
Test anxiety	41	15.9[d]	25	0	6	19	4.68[d]

Note. Different superscripts indicate significant differences between categories. [†] Each teacher could rank up to 3 items. In order to determine the average rank, non-ranked items of each teacher were replaced by expectancy values.

2.3.4 Selection and Ranking of Homework

Teachers' selection and ranking of three types of information about students' homework are displayed in Table 2-5. According to the response frequency, teachers considered all three types of information as highly valuable. Differences between the items were not statistically significant (Wilks-Lambda = .972, $F(2, 257) = 3.755$, $p = .025$, $\eta_p^2 = .028$). Finish homework on time and finish homework independently ranked higher than finish homework correctly ($\chi^2(2) = 44.425$, $p < .001$).

Table 2-5

Teachers' Selection and Ranking of Homework to Judge Student Achievement

	Selection		Ranking				
	n	%	*n*	First	Second	Third	Average rank[†]
Finish on time	258	99.6	248	85	120	43	1.82[a]
Finish independently	253	97.7	245	113	62	70	1.83[a]
Finish correctly	253	97.7	243	47	65	131	2.35[b]

Note. Different superscripts indicate significant differences between categories. [†] Each teacher could rank up to 3 items. In order to determine the average rank, non-ranked items of each teacher were replaced by expectancy values.

2.3.5 Selection and Ranking of Behavior after Class

The results of the one-way repeated measures ANOVA (Wilks-Lambda = .226, $F(4, 254) = 217.484$, $p < .001$, $\eta_p^2 = .774$) indicate statistically significant differences among the selection of the five types of information. As can be seen in Table 2-6, teachers listed specific student behaviors, including like to ask questions and like to talk with teachers after class, as main sources for judging students' achievement. However, teachers when reflecting their cognitive strategies for judging student achievement were less likely to recognize attending tutoring as required or useful after-class information.

The ranking of different after-class behavior was further measured and teachers' significantly diverging perceptions of the importance ($\chi^2(4) = 509.785$, $p < .001$) were consistent with their selection. The mean rank of questioning and talking with teachers after class were significantly higher than the rank of attending competitions and helping teachers. Attending tutoring got the lowest rank.

Table 2-6

Teachers' Selection and Ranking of Behavior after Class to Judge Student Achievement

	Selection		Ranking				
	n	%	*n*	First	Second	Third	Average rank[†]
Likes to ask you questions	245	94.6[a]	234	125	77	32	1.79[a]
Likes to talk with you	245	94.6[a]	231	96	98	37	1.96[a]
Attends competitions	124	47.9[b]	93	9	19	65	3.70[b]
Likes to help you	121	46.7[b]	105	11	31	63	3.57[b]
Attends tutoring	76	29.3[c]	54	2	18	34	3.98[c]

Note. Different superscripts indicate significant differences between categories. [†] Each teacher could rank up to 3 items. In order to determine the average rank, non-ranked items of each teacher were replaced by expectancy values.

2.3.6 Selection and Ranking of Student Demographics

Teachers were asked to list their selection and ranking of the different types of demographic information about students. Results of the one-way repeated measures ANOVA (Wilks-Lambda = .268, $F(4, 255) = 174.417$, $p < .001$, $\eta_p^2 = .732$) indicate statistically significant differences among the five types of information. It can be seen in Table 2-7 that parents' education level was utilized the most for judging student achievement, followed by statements referring to students' age. Information about students' physical appearance was of little consequence for the evaluation process.

Rank orders document that information which was selected more often was also ranked in a higher position. The Friedman-test points to significant differences between the items of this category ($\chi^2(4) = 451.397$, $p < .001$). Parents' educational level was recognized as the most impactful information for predicting student achievement, while students' age was ranked in second position and physical appearance got the lowest rank.

Table 2-7

Teachers' Selection and Ranking of Demographics to Judge Student Achievement

	Selection		Ranking				
	n	%	*n*	First	Second	Third	Average rank[†]
Parents' educational level	229	94.6[a]	219	186	25	8	1.51[a]
Age	165	68.2[b]	150	23	75	52	2.88[b]
Parents' economic status	125	51.7[c]	113	2	62	49	3.28[c]
Gender	110	45.5[c]	99	11	39	49	3.36[c]
Physical appearance	34	14.0[d]	27	4	6	17	3.96[d]

Note. Different superscripts indicate significant differences between categories. [†] Each teacher could rank up to 3 items. In order to determine the average rank, non-ranked items of each teacher were replaced by expectancy values.

2.3.7 Selection and Ranking of Social Interactions

Among the three types of information (see Table 2-8), teachers were strongly influenced by conversations with parents and other teachers for making their predictions. However, the selection of the use of social media was significantly lower than the other two categories (Wilks-Lambda = .752, $F(2, 257) = 42.469$, $p < .001$, $\eta_p^2 = .248$). Rank order results ($\chi^2(2) = 285.442$, $p < .001$) resemble teachers' selection of information in this category.

Table 2-8

Teachers' Selection and Ranking of Social Interactions to Judge Student Achievement

	Selection		Ranking				
	n	%	*n*	First	Second	Third	Average rank[†]
Conversations with parents	248	96.9[a]	236	114	112	10	1.60[a]
Conversations with other teachers	244	95.3[a]	235	125	98	12	1.58[a]
Use of social media	185	72.3[b]	190	1	20	169	2.82[b]

Note. Different superscripts indicate significant differences between categories. [†] Each teacher could rank up to 3 items. In order to determine the average rank, non-ranked items of each teacher were replaced by expectancy values.

2.3.8 Ranking of Different Data Sources

For further comparison, teachers were asked to rank the seven data sources according to their importance in the judgment process. The average rank differences between the data sources were tested on significance by the Friedman-test ($\chi^2(6) = 1077.147$, $p < .001$). The mean rank of each data source is presented in Table 2-9. The analyses reveal that teachers mainly rely on information about students' abilities and attitudes. The majority of teachers (181 out of 250) put them in the first position. Students' behavior during class ranks in the second position with a stronger influence on teachers than other categories. Homework and tests data almost equally influence the formation of teacher judgment, followed by students' behavior after class. The impact of social interactions and students' demographics just play a minor role.

Table 2-9

Teachers' Ranking of Different Data Sources

		Ranking							
	n	First	Second	Third	Fourth	Fifth	Sixth	Seventh	Average rank
Abilities and attitudes	250	181	43	6	11	8	0	1	1.60[a]
Behavior during class	251	39	157	43	10	2	0	0	2.18[b]
Homework	239	3	14	115	79	18	8	2	3.63[c]
Tests	244	23	32	40	64	54	20	11	3.86[c]
Behavior after class	239	3	3	35	59	121	16	2	4.48[d]
Social interactions	230	0	1	7	8	18	108	88	6.02[e]
Demographics	228	1	1	3	7	13	79	124	6.22[e]

Note. Different superscripts indicate significant differences between categories.

2.4 Discussion

2.4.1 Possible Explanations and Response

The present study classified the sources of student information for making accurate teacher judgments into seven domains. To enable teachers to select from the broadest range, student cues of each information source were taken from both the findings of 34 meta-analyses and data of qualitative teacher interviews (Zhu, 2014). Thereby, the research question could be addressed of how teachers select and use information from a large variety of student data to make fair judgments about student achievement.

Taken together, the findings suggest that teachers base their judgments to varying degrees on diverse student cues to arrive at their performance ratings, rather than solely relying on any specific domain. Teachers' frequently reported cues to be more accurate were students' general intelligence; students' interest; students' learning strategies; students' engagement during class (concentrates well); independent completion of homework; and students' last test performance. These results are in line with conclusions from prior research (Haigh et al., 2013; Kishor, 1989; Oudman et al., 2018).

Furthermore, teachers could develop a clear hierarchy of data sources to judge student achievement while miscellaneous cues were available to them. The best information was rearing from student abilities and attitudes and the least important information from student demographics and social interactions of teachers. Earlier research has indicated that teachers were often influenced by non-cognitive student characteristics such as observed participation and expressed motivation (Wijnia et al., 2016). However, the present study further shows teachers' strategic preferences for both students' cognitive and non-cognitive characteristics. The perception of students' general intelligence was taken as the most significant cue for judging student achievement, followed by students' interest and learning strategies, although the differences are not significant. It seems that interviewed teachers in the qualitative studies are more likely to describe some specific situations or observed behaviors when reporting

their judgment process, as the daily interactions between teachers and students provide them with a rich picture of students' learning status (Meissel, Meyer, Yao, & Rubie-Davies, 2017). In addition, the current study reveals that students' demographic cues, specifically physical appearance, parents' economic status, and gender were found to draw the least awareness of teachers. The results suggests that—although there may be broad agreement about gender and SES biases in teacher judgments (Baudson et al., 2016; Dusek & Joseph, 1983; Holder & Kessels, 2017)—teachers intentionally try to make the best use of other student resources and bypass the stereotype of physical characteristics and economic status.

Teachers believed that they are supported by effective judgment strategies. For each source of information, teachers tend to mark the cue they mainly relied on as the most important, whereas the least marked cue got the lowest rank. In other words, teachers believe that the student cues they use most often are the most important ones that help them to develop an accurate judgment on student achievement. Comparable conclusions were drawn by Praetorius, Berner, Zeinz, Scheunpflug, and Dresel (2013) who found that the majority of teachers were overconfident of their judgments.

Moreover, it can be concluded that, in general, teachers were using student cues of high validity to generate their judgments. From the perspective of Social Judgment Theory, teachers should be informed about more valid indicators of student achievement to make more precise predictions. Student cues like general intelligence, interest, or engagement were mainly utilized to substantiate teachers' judgments. These cues are significantly correlated with students' academic achievement (Lei, Cui, & Zhou, 2018; Roth et al., 2015; Schiefele, Krapp, & Schreyer, 1993), which indicates from the perspective of the lens model a more accurate teacher judgment. However, some student cues like information from homework were regarded as much more important than covered by the literature (Cooper et al., 2006; Kim & Seo, 2015). This is in agreement with some other studies (e.g., Praetorius, Koch,

Scheunpflug, & Zeinz, 2017), which found that teachers partly use invalid sources for their judgments. It could be one of the reasons that lead to inaccurate teacher judgments.

2.4.2 Limitations and Future Research

There are several limitations to this investigation, which also suggest directions for future research. In the current study, teachers were provided with various student cues in a semi-structured questionnaire and were required to identify the cues that suit them. Even each semi-open question contained a further item "other" in case important student cues were missing, teachers seldom made use of this option. The pre-structured questionnaire somehow restrains teachers from thinking outside the box.

According to the framework of Social Judgment Theory, the judgment process includes three different paths. Sufficient studies on two paths have been conducted so that meta-analytic findings exist on the accuracy of teacher judgment (Hoge & Coladarci, 1989; Südkamp et al., 2012) as well as on the validity of student information cues (see Table 2-1). We therefore only focused on teachers' utilization of different types of student cues. However, it is still interesting to know whether teachers of the current sample can accurately judge student achievement. Results of prior research have shown that Chinese primary school teachers can be very accurate judges (Zhu & Urhahne, 2015). However, to gain insight into the cue utilization validities, further studies might examine teachers' cue utilization and judgment accuracy simultaneously. Furthermore, the study tried to explain the validity of student cues with the results from 34 meta-analyses. Even though the meta-analyses could give some valuable hints, it is necessary to conduct studies testing cue validity and cue utilization concurrently with the same samples. Thus, an integrated analysis of all three paths of the Lens model should be a goal of future research.

2.4.3 Conclusions

By using the lens model of Social Judgment Theory, this study has confirmed the path of cue utilization in the judgment process. Teacher obtained students cues for their

achievement judgments from seven information sources. The most frequently used cues are general intelligence, interest, and learning strategies in the source of abilities and attitudes; concentration during class; finishing homework punctually, independently, and correctly; last test performance; questioning and talking with teachers after class; conversations with parents and colleagues, and parents' educational level. Teachers believed that the student cues they were using are predictive and of value. Particularly, they considered student cues from abilities and attitudes, behavior during class, homework, and tests as most valid for their judgments.

Most of the student cues that teachers perceive as important are in reality associated with students' actual achievement. Yet, teachers may overestimate the validities of some cues. Therefore, they should be informed what kind of student information can be regarded as trustworthy and helps them to arrive at fair judgments of student achievement.

CHAPTER 3: THE USE OF LEARNER RESPONSE SYSTEMS IN THE CLASSROOM ENHANCES TEACHERS' JUDGMENT ACCURACY

The second study of the dissertation has been published in the journal of Learning and Instruction. It is available as an online version and in printed form:

Zhu, C., & Urhahne, D. (2018). The use of learner response systems in the classroom enhances teachers' judgment accuracy. *Learning and Instruction, 58*, 255–262. doi: 10.1016/j.learninstruc.2018.07.011

A brief summary of the study is presented below.

3.1 Summary of Study 2

Different ways have been discussed to improve teacher judgment accuracy such as systemic training programs (Trittel, Gerich, & Schmitz, 2014), professional development (Thiede et al., 2015), self-monitoring with diaries (Klug, Gerich, Bruder, & Schmitz, 2012), or simulated classrooms (Südkamp & Praetorius, 2017). With the use of learner response systems (clickers) in the classroom, we tried to provide teachers with more feedback about student achievement. By checking learning protocols, teachers should come to know the difficulty of tasks as well as the difficulties of students. Teachers will enhance judgment accuracy when they know more about their students and can offer them better suited learning tasks, which may result in higher student achievement.

A pretest-posttest-intervention study with one control and two experimental groups was conducted in German middle school over a period of five weeks. Nineteen classes (5 control, 7 diary, and 7 clickers) with a total of 428 sixth-grade students and 18 mathematics teachers took part in the investigation. Students worked on a standardized mathematics test (DEMAT6+; Götz, Lingel, & Schneider, 2013), while teachers estimated student test scores. Results show a significant improvement of all three groups in mathematics achievement from pre- to post-test. However, clicker groups had significantly higher learning gains than the other two groups, which increased test performance in a similar manner. On the first point of measurement, teachers were moderate judges of student achievement but improved significantly in the post-test. Teachers in the control group did not change significantly, teachers of the diary group to a lower and teachers in the clicker group to a higher extent. The use of clickers in the classroom is a time-saving and efficient way to enhance student achievement and teacher judgment accuracy.

CHAPTER 4: TEMPORAL STABILITY OF TEACHERS' JUDGMENT ACCURACY OF STUDENTS' MOTIVATION, EMOTION AND ACHIEVEMENT

4.1 Introduction

Every student wants to be judged fairly and every teacher wants to be fair (Dalbert, Schneidewind, & Saalbach, 2007). In order to judge fairly, the accuracy of teacher judgment has to be highly reliable (Meissel, Mayer, Yao, & Rubie-Davies, 2017). Given high reliability, teachers are able to make fair judgments at different times. Reliability of the accuracy of teacher judgment can be determined by the test-retest method. Teacher judgments and student characteristics are measured twice within a short timeframe. Comparing the relationships of both measures provides information about the temporal stability of teacher judgment accuracy.

In the literature, however, there are hardly any studies that deal with teachers' judgment accuracy over time. This may be due to the fact that educational researchers in the past have been interested in the opposite issue. Teacher expectancy research focused on the question of how students develop when teacher judgment is not precise but inaccurate (e.g., Brophy, 1983; Friedrich, Flunger, Nagengast, Jonkmann, & Trautwein, 2015; Hinnant, O'Brien, & Ghazarian, 2009; Jamil, Larsen, & Hamre, 2018; Zhu, Urhahne, & Rubie-Davies, 2018). In order to generate expectancy effects, subjective teacher judgments could not correspond with actual student achievement or real student attributes. As it was important that teacher assessments had to be inaccurate, their measurement reliability or temporal stability was not closely investigated.

When the stability of teacher judgment accuracy was further examined, this usually happened over a longer period of time. Often considered were periods of half a year or more (Hinnant et al., 2009; Lorenz & Artelt, 2009; Oerke, McElvany, Ohle, Ullrich, & Horz, 2016; Stang & Urhahne, 2016). The aim was to show that the accuracy of teacher judgment stays on a high level or even increases through longer experiences with the class (Oerke et al., 2016). The problem with these studies is that the reasons for temporal stability of teacher judgment accuracy are not quite clear. If the accuracy of teacher judgment is time-stable or even increasing, this outcome may not necessarily be due to teachers' diagnostic competence. It

could also be that the students—in the sense of a self-fulfilling prophecy—adapted themselves to teachers' judgments, as has often been shown by teacher expectancy research (e.g., Jussim & Eccles, 1995; Jussim & Harber, 2005; Rosenthal, 1991; Rosenthal & Jacobson, 1968).

It is therefore advisable to consider a shorter period of time in order to determine the temporal stability of teacher judgment accuracy. This would ensure that student characteristics hardly change or do not change at all so that accurate judgments really depend on teachers' diagnostic competence. The period between surveys should be short but long enough that teachers could not simply remember their judgments from the last questionnaire. This is typically given after a time interval of four weeks (Rammstedt & Rammsayer, 2002).

Moreover, teachers should not only be able to assess single student characteristics well, but properly judge students as a whole (Huber & Seidel, 2018). To this end, teachers should repeatedly rely in their judgments on the same indicators. In this study, the accuracy of teacher judgment is examined in terms of motivational, emotional, and cognitive student characteristics. The correspondence between teacher judgments and student characteristics is studied at two different points of time in order to determine the accuracy of teacher judgment over time. Through use of structural equation modeling and measurement invariance testing, it is further explored whether teachers' judgments are repeatedly based on the same motivational, emotional, and cognitive student characteristics.

4.1.1 Teacher Judgment Accuracy of Students' Motivation, Emotion, and Achievement

The accuracy of teacher judgment has been intensively studied so that concrete expectations can be formulated about the size of the relationship between teacher judgments and student characteristics. High correlations larger than .60 are typically found in the cognitive domain (Hoge & Coladarci, 1989; Südkamp et al., 2012). Moderate correlation between .30 and .60 to small correlations of less than .30 are the rule when teachers try to accurately judge students' motivation and emotion (Urhahne & Zhu, 2015).

In a meta-analysis, Hoge and Coladarci (1989) found that teacher judgments of student achievement showed a median correlation of .66 with actual student achievement. Two decades later, Südkamp et al. (2012) repeated the meta-analysis on an updated data basis and came to very similar results. A mean correlation of .63 indicated a high relationship between teacher judgment and student achievement. In general, teachers possess a solid foundation for making fair judgments about student achievement and decisions about school careers.

Besides, teachers are responsible to properly assess students' motivation and emotion (Dicke, Lüdtke, Trautwein, Nagy, & Nagy, 2012). They should at least tentatively know whether their students are self-confident and willing to make an effort, study with interest, or anxiously look forward to the upcoming exams. Teachers can judge those motivational variables well that are closely linked to teachers' grading. Various studies have shown that teachers can predict students' expectancy for success in the next exam with high accuracy and level of aspiration for the next exam with moderate accuracy (Urhahne et al., 2011; Urhahne, Timm, Zhu, & Tang, 2013; Urhahne et al., 2010). Somewhat more difficult is to correctly judge students' academic self-concept and self-efficacy, often resulting in moderate correlations (Givvin, Stipek, Salmon, & MacGyvers, 2001; Helm, Müller-Kalthoff, Mukowski, & Möller, 2018; Praetorius, Berner, Zeinz, Scheunpflug, & Dresel, 2013; Spinath, 2005). Both hypothetical constructs are good predictors of academic achievement (Lee & Stankov, 2018), but not congruent with it. There are probably students with high achievement, but low self-concept and low self-efficacy as well as students for whom this ratio is reversed. Moreover, teachers are doing hard to judge students' learning effort due to its reciprocal relationship with student ability. If two students perform equally well, the teacher will find the one more capable of doing so with less effort (Frieze & Weiner, 1971; Weiner, 1980). Since the interplay of ability and effort in student achievement is often unclear, teachers' judgment accuracy is only on a moderate level (Urhahne et al., 2010; Wright & Wiese, 1988; Zhu & Urhahne, 2014). In addition to motivational factors, emotions are important for learning at

school (Pekrun, Muis, Frenzel, & Goetz, 2018). Individual interest is both a way and a goal of learning. Interested students achieve higher learning outcomes (Schiefele, Krapp, & Winteler, 1992). For teachers, interest is an essential variable as it can be specifically stimulated in the classroom (Bergin, 1999). Teachers' judgments of student interest, however, reveal little more than small correlations with student data (Karing, 2009; Zhu & Urhahne, 2014). Test anxiety is the best-studied learning emotion (Zeidner, 1998). When teachers assess students' test anxiety, often only small correlations to student self-report data can be found (Karing, Dörfler, & Artelt, 2015; Spinath, 2005).

The Realistic Accuracy Model of Funder (2012) provides a good explanation for different teacher judgment accuracy of cognitive, motivational and emotional variables. In order to correctly assess a student characteristic, the student must make (a) behavior available that provides (b) relevant information for judging the hypothetical construct fairly. The teacher has to (c) detect the behavior and (d) utilize the information correctly by drawing the right inferences. Only when these four conditions are met, teachers can make accurate judgments of student characteristics. Student achievement is much easier to judge as a bunch of relevant information such as quality of homework, verbal contributions in class, or written exams is permanently available. Judgments of motivation and emotion might be much more complicated as teachers need to look for suitable indicators in student behavior.

4.1.2 Stability of Teacher Judgment Accuracy

Longitudinal studies to measure the stability of teacher judgment accuracy are almost exclusively related to student achievement. Only the study by Givvin et al. (2001) provides some information on the temporal stability of teacher judgment accuracy on student motivation and emotion.

In a study by Oerke et al. (2016), teachers were required to rate students' ability of text-picture integration after half a year and one and a half years of contact. Teachers tended

to overestimate student achievement and had moderate accuracy on the rank component. Level and rank component did not change significantly within one year.

Lorenz and Artelt (2009) examined the diagnostic skills of elementary school teachers within a time interval of six months. In the areas of vocabulary, text comprehension and arithmetics, teachers were able to predict student achievement on standardized tests with moderate accuracy at both times. Rank component differences between measurement points were not significant.

Paleczek, Seifert, and Gasteiger-Klicpera (2017) provided similar findings with a study on reading abilities of second and third grade students. The two student groups were tested for decoding and reading skills at the beginning and end of the school year. In both grades, moderate accuracy of teacher judgment did not change significantly during the school year.

The longitudinal study by Hinnant et al. (2009) covered a period of four years. Teachers were asked to rate reading and mathematics abilities in the first, third, and fifth grade. Moderate correlations in both subjects were found at all times, which did not deviate significantly from each other.

Stang and Urhahne (2016) examined judgment accuracy of secondary school teachers in mathematics twice within a time period of six months. They found significant improvements in the rank component and the level component. However, teachers' predictions of test results at the first point of measurement were not very accurate. The differentiation component was unaffected by changes.

Givvin et al. (2001) asked teachers to rate motivation and emotion of selected students four times within a school year. The first and last time of measurement can be compared with each other as they both refer to mathematics in general. Teacher judgment accuracy on perceived ability and learning orientation did not change over the school year but stayed at a low level. The accuracy of teacher judgment was even lower for positive and negative

emotions. However, teacher judgment accuracy of negative emotions improved significantly from low to moderate degree at the end of the year.

To sum up, previous studies on the accuracy of teacher judgment point to a relatively high degree of temporal stability. Teacher judgment rarely becomes more accurate over time (e.g., Givvin et al., 2001; Stang & Urhahne, 2016). The correlations between teacher ratings and student characteristics are strongest in the cognitive domain and weakest in the emotional area.

For this study, it can be hypothesized that the accuracy of teacher judgment changes only slightly or not at all over a short period of four weeks, and best teacher ratings may occur on achievement-related variables (Funder, 2012). In addition, this study examines the interplay of various motivational, emotional, and cognitive factors to holistically explore the accuracy of teacher judgments of student characteristics (Huber & Seidel, 2018).

4.2 Method

4.2.1 Participants

The sample consisted of nine classes from a Chinese elementary school with class sizes ranging from 32 to 42 students. The 326 sixth-graders were between 10 and 14 years old (M = 11.79, SD = .69), including 148 girls and 177 boys (1 missing gender). The students were taught by five mathematics teachers aged 27 to 45 years (M = 34.80, SD = 8.38), with four of the five teachers responsible for two classes. The all-female teachers had an average teaching experience of M = 12.40 years (SD = 8.14) and already knew the classes since 3 to 5 years (M = 4.20, SD = 1.10).

4.2.2 Materials

Mathematics achievement. A standardized mathematics test for the sixth grade (DEMAT 6+, Götz, Lingel, & Schneider, 2013) was used to measure students' mathematics achievement. A native speaker translated the test items into Chinese. Teachers were asked to make sure that the test corresponds to the Chinese mathematics curriculum. The DEMAT 6+

encompasses 35 tasks and is valid at the end of the school year. As the study took part in the first half of sixth grade, the test was reduced to those 25 items especially relevant in the first half year. Cronbach's alpha, as a measure of the internal consistency, amounted to $\alpha_{t1} = .78$ in the first test and $\alpha_{t2} = .81$ in the second test.

Motivation. In order to obtain a comprehensive picture of students' motivation, five different indicators were selected from the Ulm Motivational Test Battery (Ziegler, Dresel, Schober, & Stöger, 2005; Ziegler, Dresel, & Stöger, 2008). The adopted scales have already been translated into Chinese and successfully applied in previous studies (Urhahne et al., 2010; Zhou & Urhahne, 2013; Zhu & Urhahne, 2014, 2015). Expectancy for success ('What do you think: What grade will you get on your next mathematics test?') and level of aspiration ('What is the minimum grade on your next mathematics test that you would be satisfied with?') were queried with one item each with respect to the next mathematics test. Students should specify a score on the Chinese grading scale from 0 to 100. Self-efficacy beliefs (e.g., 'When a problem in mathematics arises, I can master it on my own', $\alpha_{t1} = .82$ resp. $\alpha_{t2} = .86$), learning effort (e.g., 'I do my best in mathematics', $\alpha_{t1} = .76$ resp. $\alpha_{t2} = .77$) and academic self-concept (e.g., 'I am good in mathematics', $\alpha_{t1} = .88$ resp. $\alpha_{t2} = .88$) were rated on five-point Likert scales from 1 – '*not at all true*' to 5 – '*very true*' with six items each with satisfactory reliability at both times of measurement.

Emotion. Test anxiety and interest were measured with scales according to the Chinese version of the Academic Emotions Questionnaire-Mathematics (AEQ-M; Pekrun, Frenzel, Goetz, & He, 2005). The six items each were rated in the same way as the motivation items. The reliability for individual interest in the subject of mathematics (e.g., 'I am looking forward to mathematics lessons', $\alpha_{t1} = .88$ resp. $\alpha_{t2} = .88$) was satisfactory at both times of measurement. In case of test anxiety (e.g., 'I am afraid to get bad grades in mathematics', α_{t1}

= .69 resp. α_{t2} = .64), measurement reliability was slightly reduced and fell below the desired value of .70.

Teacher materials. Mathematics teachers received a copy of the standardized test that was used to measure students' mathematics achievement. The teachers assessed for each student in class the number of correctly solved tasks in the standardized test. In addition, they were requested to judge for each student the motivational and emotional constructs measured by scales in comparison to other students of the same age on nine-point Likert scales (1 = *extremely low*, 5 = *average*, 9 = *extremely high*). For comparative analyses of teacher and student data, teachers' judgments were later transformed to a five-point rating scale. Expectancy for success and level of aspiration were measured differently: teachers had to predict the score that students would expect to get and would be satisfied with in the next mathematics test on the Chinese grading scale from 0 to 100.

4.2.3 Procedure

The school principal was asked for permission to conduct the research study at the elementary school. Students' parents were informed by the mathematics teachers that two surveys will be distributed to the students within a timeframe of four weeks. All students were given parental permission to participate in the investigation. At the beginning of the lesson, trained investigators carried out the standardized mathematics test. Afterwards, students were asked to fill in the questionnaire items and scales on motivation and emotion. The mathematics teachers supported the completion of the survey, which could have been finished within one lesson (40 minutes). In the afternoon, the teachers themselves had the opportunity to assess test performance and motivational-affective characteristics of each student in class. After four weeks, the same procedure was repeated with the identical students and teachers.

4.2.4 Statistical Analyses

Stability of the rank component was checked by testing correlation differences between the two points of measurement at the individual level. Stability of the level

component and the differentiation component were tested by means of dependent t-tests at the class level.

The data set has a multi-level structure in which the participating sixth-graders are nested into classes (Garson, 2013). Student characteristics of the same class may be more similar to each other than student characteristics of other classes. In order to rule out that classes greatly differ and teachers had to judge very different classes, the intra-class correlations (ICC1) were computed for all variables of the first and second point of measurement. It turned out that the classes involved were relatively similar to each other. In the mathematics test, e.g., only 3.4% of the variance at t_1 and 4.2% of the variance at t_2 was due to achievement differences between classes. The differences in the motivational-affective characteristics between classes varied between 0.5% (test anxiety) and 8.1% (expectancy for success) at the first point of measurement. At the second point of measurement, the values were slightly higher and varied between 0.9% (test anxiety) and 11.5% (expectancy for success). In view of the small differences between classes and the insufficient number of classes, the multi-level structure was not taken into account.

Further analyses should show to what extent teacher judgment accuracy remains temporally stable not only over individual constructs, but over all motivational and emotional student characteristics. Structural equation models based on manifest variables were computed using AMOS 25 (Arbuckle, 2017). Manifest instead of latent variables were computed as all teacher judgments and two student motivation variables were measured with just one item each. Missing values in the data were replaced by the expectation-maximization method, which allows valid maximum likelihood estimates for means, variances, and covariances (Allison, 2002; Schafer & Graham, 2002). At the first point of measurement up to 3.7% (effort) and at the second point of measurement up to 4.3% (test anxiety) of the student data had to be estimated.

To determine the model fit of the structural equation models, various goodness-of-fit indices were taken into account. The comparative fit index (CFI), the Tucker-Lewis index (TLI), and the root mean square error of approximation (RMSEA) are widely used evaluation criteria in structural equation modeling. CFI and TLI values above .95 represent a good model fit (Hu & Bentler, 1999). RMSEA values should below .08 for an adequate fit, or below .06 for a good model fit (Hu & Bentler, 1999).

Measurement invariance was tested to make a comparison between models of the first and second point of measurement. Configural invariance deals with the issue whether the measurement model remains constant at both times of measurement. This would suggest that teachers' judgment accuracy relies at both times on the same indicators. Configurational invariance is also referred to as pattern invariance and represents a baseline model for further comparisons. Metric invariance builds on configural invariance and asks whether the factor loadings of the constructs remain constant over the times of measurement. This would suggest that the accuracy of teacher judgment over time is similarly influenced by the same motivational and emotional constructs.

Finally, in order to test the invariance models against each other and compare more and less restrictive models, Cheung and Rensvold (2002) as well as Chen (2007) have provided certain recommendations. When testing for measurement invariance and the sample size is sufficient ($N > 300$), goodness-of-fit indices CFI and TLI should not drop more than 0.01 and RMSEA should not increase more than 0.015. If these conditions are met, metric (factor loadings) invariance can be assumed.

4.3 Results

4.3.1 Stability of Student Characteristics and Teacher Judgments

In a first step, repeated measures analyses of variance were used to check whether student characteristics and teacher judgments changed over time. Table 4-1 shows that student characteristics did not significantly change in the short four-week period. This is an important

prerequisite for evaluating the accuracy of teacher judgment as students did not simply adapt to teachers' former judgments. Especially striking is the high performance of Chinese students in the mathematics test: on average, 20 out of 25 items were solved correctly. Consequentially, students show high expectancy for success and high level of aspiration. Learning effort is the highest rated motivational-affective variable among the Chinese students. Teacher judgments in Table 4-1 reveal four significant changes over time. Teachers reduced their high expectation of student achievement from an average of 21 to 20 correctly solved items. Furthermore, teachers assumed higher student self-concept, increased expectancy for success and reduced test anxiety at the second point of measurement.

Table 4-1

Stability of Student Characteristics and Teacher Judgments

	First time of measurement		Second time of measurement				
	M	*SD*	*M*	*SD*	*F*	*df*	η^2
Student characteristics							
Achievement	19.94	3.77	20.25	3.40	.99	325	.003
Self-concept	3.34	.98	3.34	1.01	.00	325	.000
Self-efficacy	3.44	.88	3.44	.94	.01	325	.000
Effort	4.20	.68	4.19	.68	.06	325	.000
Expectancy for success	89.44	10.76	90.91	11.45	2.93	325	.009
Level of aspiration	94.32	7.34	94.82	8.40	.62	325	.002
Interest	3.97	.88	3.91	.91	.65	325	.002
Test anxiety	3.31	.88	3.33	.65	.20	325	.001
Teacher judgments							
Achievement	21.05	3.78	20.06	3.85	10.62	325	.032***
Self-concept	6.83	1.53	7.14	1.44	6.99	325	.021**
Self-efficacy	7.61	1.38	7.45	1.33	2.12	325	.006
Effort	6.98	1.56	7.14	1.42	1.96	325	.006
Expectancy for success	89.07	12.28	91.15	10.12	5.86	325	.018*
Level of aspiration	93.77	9.19	93.87	8.45	.03	325	.000
Interest	7.22	1.52	7.29	1.44	.30	325	.001
Test anxiety	2.69	1.55	2.33	1.32	10.01	325	.030**

Note. *** $p < .001$; ** $p < .01$; * $p < .05$.

4.3.2 Stability of Teacher Judgment Accuracy for Single Variables

In a second step, teacher judgments and student characteristics were related to each other. Table 4-2 shows the temporal stability of the three components of judgment accuracy. The results of rank component are in line with the hypotheses. Teachers are able to judge student achievement with high accuracy, motivational student characteristics with moderate to high accuracy, and emotional student characteristics mostly with low accuracy. Values of the rank component remain largely constant over time. Only teachers' judgments of students' self-concept gets worse from first to second measurement. The level component in Table 4-2 reveals in general small differences between teacher judgment and student self-report. Teachers, however, greatly underestimate students' test anxiety. This judgment tendency is even stronger at the second point of measurement. The differentiation component hardly changes over time. Significant changes only occur with respect to student achievement. The heterogeneity of student achievement is judged more badly at the second point of measurement.

Table 4-2

Stability of Teachers' Judgment Accuracy of Students' Achievement, Motivation and Emotion

	First time of measurement		Second time of measurement			
	M	*SD*	*M*	*SD*	*z* resp. *t*	*df*
Rank component						
Achievement	.70	.22	.74	.17	-1.06	325
Self-concept	.65	.11	.54	.10	2.18*	325
Self-efficacy	.52	.14	.47	.13	0.84	325
Effort	.42	.19	.42	.11	0.00	325
Expectancy for success	.65	.14	.61	.13	0.84	325
Level of aspiration	.56	.17	.51	.19	0.89	325
Interest	.33	.20	.35	.14	-0.29	325
Test anxiety	.28	.17	.16	.17	1.61	325
Level component						

Achievement	1.09	1.09	-.10	1.68	1.57	8
Self-concept	.58	.29	.73	.50	-1.64	8
Self-efficacy	.86	.40	.81	.37	.50	8
Effort	-.20	.19	-.11	.26	-1.88	8
Expectancy for success	-.16	8.38	.31	6.70	-.54	8
Level of aspiration	-.61	4.84	-1.08	5.19	.91	8
Interest	.14	.23	.24	.20	-2.04	8
Test anxiety	-1.43	.39	-1.70	.31	3.14*	8
Differentiation component						
Achievement	0.99	0.21	0.86	0.21	2.67*	8
Self-concept	0.77	0.14	0.64	0.18	2.18	8
Self-efficacy	0.76	0.24	0.70	0.25	.44	8
Effort	1.18	0.33	1.03	0.15	1.43	8
Expectancy for success	1.07	0.69	0.95	0.50	1.29	8
Level of aspiration	1.44	0.94	1.32	0.85	.62	8
Interest	0.88	0.25	0.78	0.16	1.16	8
Test anxiety	0.83	0.34	0.80	0.24	.47	8

Note. * $p < .05$.

4.3.3 Stability of Teacher Judgment Accuracy for Multiple Variables

The intercorrelations of teacher judgments and student characteristics in Table 4-3 are quite strong. While all teacher variables are highly correlated, student variables correlate on a moderate to high level. Test anxiety, which is little related to other student variables, is an exception.

Table 4-3

Intercorrelations of Teacher Measures (Upper Half) and Student Measures (Lower Half)

	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)
(1) Achievement	—	.83/.70	.85/.77	.83/.75	.80/.76	.77/63	.85/.78	-.80/-.61
(2) Self-concept	.54/.50	—	.74/.84	.81/.87	.66/.75	.59/.64	.81/.87	-.76/-.60
(3) Self-efficacy	.41/.36	.75/.81	—	.85/.92	.70/.74	.70/.67	.89/.95	-.80/.-66
(4) Effort	.44/.39	.62/.63	.65/.64	—	.65/.75	.62/.69	.91/.95	-.74/-.66
(5) Expectancy for success	.61/.52	.66/.56	.59/.52	.57/.52	—	.93/.94	.69/.78	-.72/-.80
(6) Level of aspiration	.52/.49	.52/.43	.51/.39	.50/.42	.81/.76	—	.67/.72	-.72/-.79
(7) Interest	.39/.30	.58/.59	.54/.56	.65/.73	.48/.44	.43/.31	—	-.82/-.68
(8) Test anxiety	-.26/-.13	-.49/-.26	-.37/-.20	-.28/-.12	-.33/-.13	-.22/-.10	-.31/-.08	—

Note. First correlation in each cell indicates the first time of measurement, second correlation stands for the second time of measurement; $p < .05$ for all $r > |.12|$; $p < .01$ for all $r > |.15|$; $p < .001$ for all $r > |.19|$.

Structural equation models were computed to look at the correspondence between teacher judgments and student characteristics not only at an individual, but at a more holistic level. Figure 4-1 graphically depicts the relationships between multiple teacher judgments and multiple student characteristics. This motivation model includes all motivational and emotional variables combined to two global latent variables – one for the teachers and one for the students. As illustrated in Table 4-3, the motivation model has a good model fit at both times of measurement. CFI and TLI are higher than .95 and RMSEA is smaller than .08 (Hu & Bentler, 1999). Invariance testing shows that the model with configural measurement invariance does not significantly differ from the model with factor invariance. The changes in CFI and TLI are less than 0.01 and RMSEA does not increase more than 0.015 (Chen, 2007; Cheung & Rensvold, 2002). Thus, when comparing motivation models 1 and 2 of Table 4-4, measurement invariance over time can be assumed.

Table 4-4

Model Comparison of Structural Equation Modeling for Teachers' Judgments and Students' Characteristics

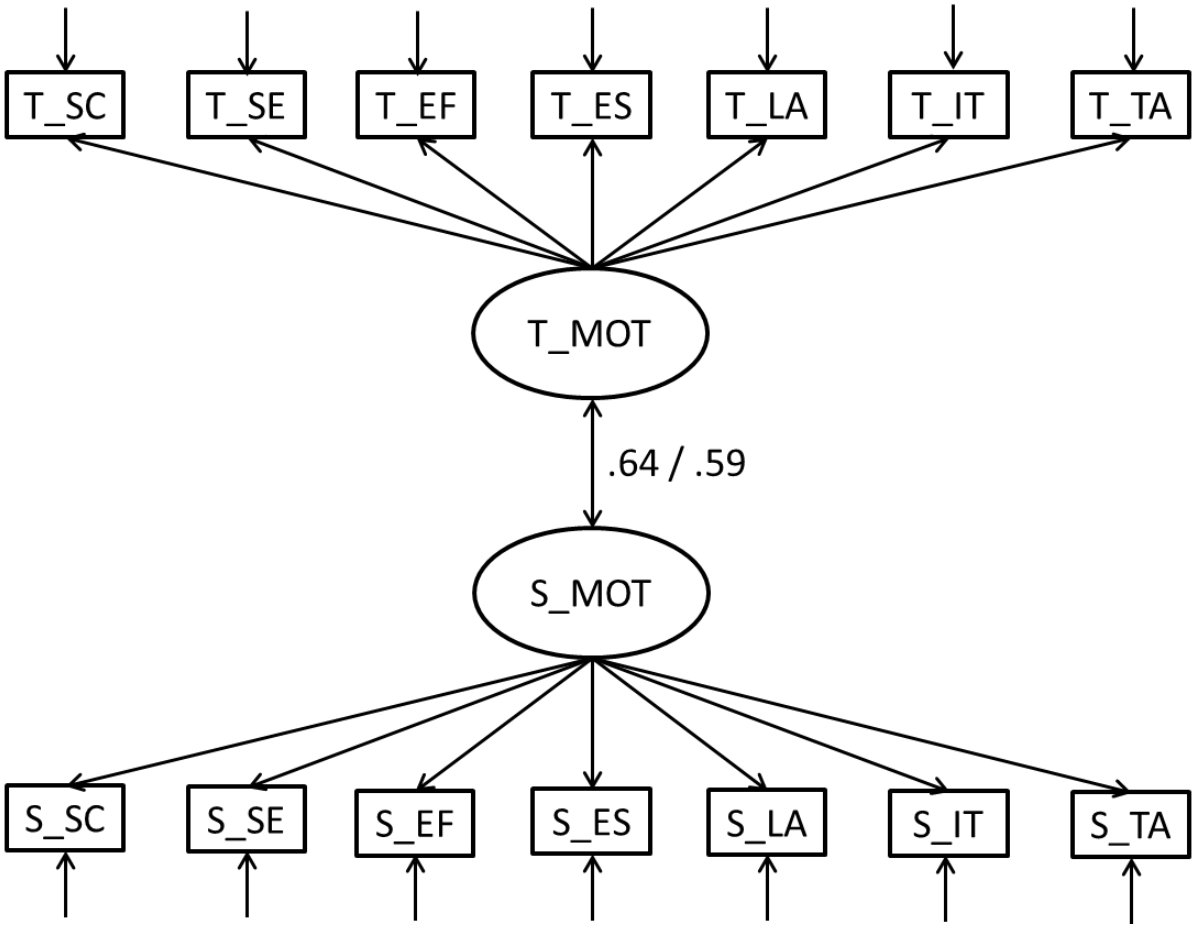
Model		χ^2	*df*	CFI	TLI	RMSEA
1	Measurement model with motivation at Time 1	108.198	49	.987	.975	.061
2	Measurement model with motivation at Time 2	113.831	47	.986	.973	.066
3	Measurement model with motivation and achievement at Time 1	164.418	72	.983	.971	.063
4	Measurement model with motivation and achievement at Time 2	196.882	69	.977	.960	.076
Invariance across Time for Models 1 & 2						
I1	Configural invariance	206.179	94	.988	.977	.043
I2	Invariance of factor loadings	241.937	108	.986	.976	.044
Invariance across Time for Models 3 & 4						
I3	Configural invariance	345.329	138	.981	.967	.048

I4	Invariance of factor loadings	393.673	155	.978	.966	.049

Note. df = degrees of freedom; CFI = comparative fit index; TLI = Tucker-Lewis index; RMSEA = root mean square error of approximation. All χ^2 values are significant ($p < .001$).

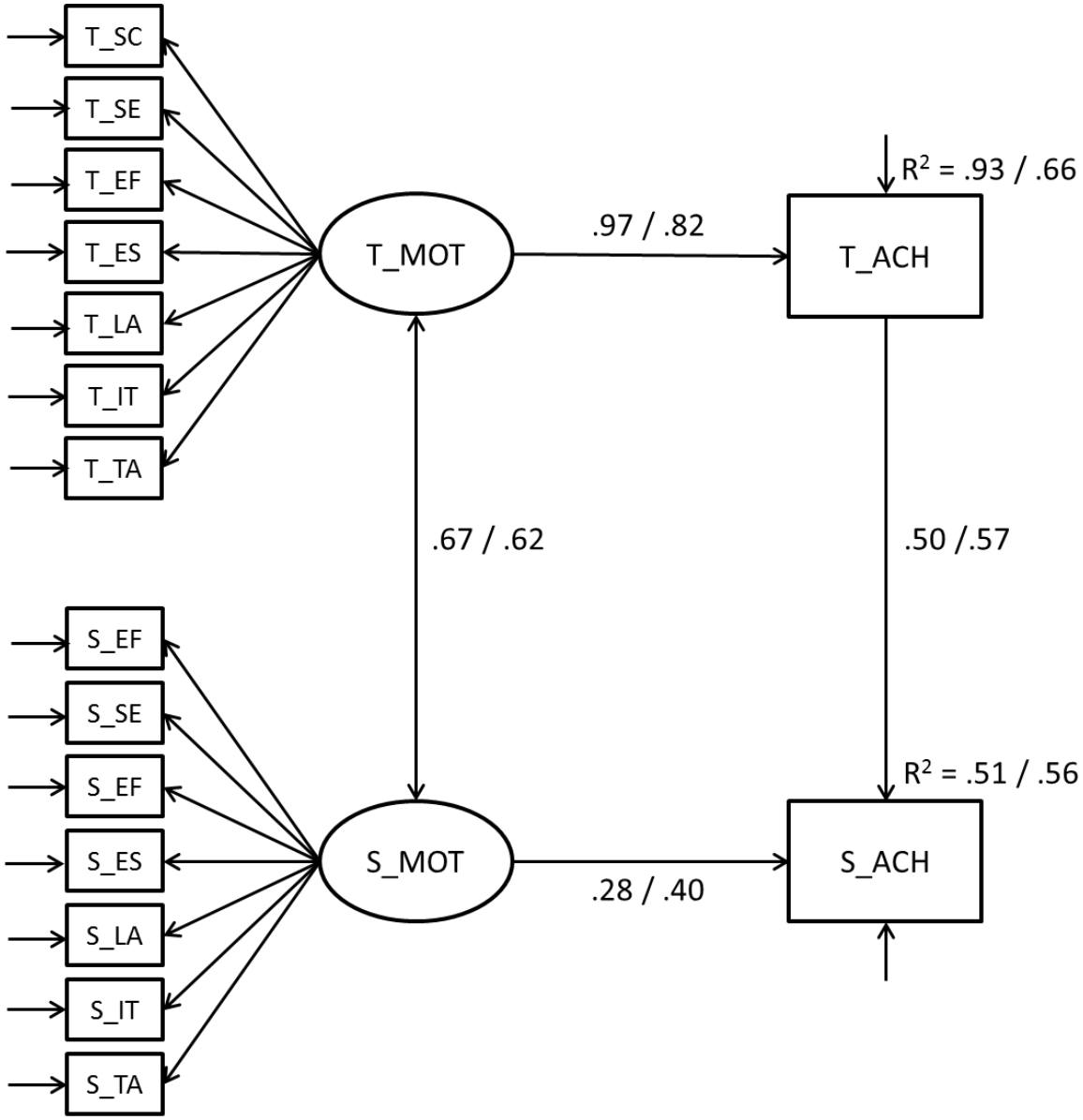
In Figure 4-2, the motivation model has been extended by the cognitive variables. It is striking that teacher judgment of students' motivation and emotion is a good predictor of teacher judgment of student achievement. On the other hand, student motivation and emotion can predict actual student achievement far less accurately. From Table 4-4, it can be seen that the measurement model with motivation and achievement shows good model fit at the first point of measurement and adequate model fit at the second point of measurement. In the invariance tests CFI, TLI and RMSEA change only slightly from the model with configural measurement invariance to the model with factor invariance. Thus, when comparing models 3 and 4 of Table 4-4, measurement invariance over time can be assumed.

Figure 4-1. Structural Equation Model on the Relations between Teachers' Judgments and Students' Motivation and Emotion



Note. The upper half of the figure shows motivational and emotional judgments of the teachers (T), while the lower half depicts motivational and emotional characteristics of the students (S). SC = self-concept, SE = self-efficacy, EF = effort, ES = expectancy for success, LA = level of aspiration, IT = interest, TA = test anxiety, MOT = motivation and emotion.

Figure 4-2. Structural Equation Model on the Relations between Teachers' Judgments and Students' Motivation and Emotion to Predict Student Achievement



Note. The upper half of the figure shows motivational, emotional and cognitive judgments of the teachers (T), while the lower half depicts motivational, emotional and cognitive characteristics of the students (S). SC = self-concept, SE = self-efficacy, EF = effort, ES = expectancy for success, LA = level of aspiration, IT = interest, TA = test anxiety, MOT = motivation and emotion, ACH = achievement.

4.4 Discussion

4.4.1 Summary of Findings

The aim of the study was to examine the temporal stability of teachers' judgment accuracy of students' motivation, emotion, and achievement. The results indicate that teachers can judge their students' achievement with high accuracy, motivation with moderate to high accuracy, and emotion mostly with low accuracy. The stability of teachers' judgment accuracy over a four-week time interval was high with little changes on the three dimensions.

4.4.2 Possible Explanations and Response

Consistent with prior research (e.g., Dicke et al., 2012; Zhu & Urhahne, 2014), we found that teachers could predict students' academic performance much better than students' academic motivation. Meanwhile, teachers had difficulties in determining students' academic emotion. To explain these findings, many researchers argue that it is more realistic for teachers to collect and analyze information about students' academic outcomes through homework, exercises, and tests. Motivational-affective traits are less stable over time and more difficult to detect and interpret (Givvin et al., 2001; Karing et al., 2015; Zhu & Urhahne, 2014). When considering the results against a cultural background, there are other factors that might have contributed to discrepancies between teachers' prediction and students' self-reflection of their learning motivation and emotions. Chinese teachers place a high value on students' learning outcomes which might explain the high accuracy of their achievement judgments. On the other hand, when these teachers have to take care of a large classroom with a high number of students, they could hardly be supportive to anyone and be strongly involved with their students' emotionality.

Furthermore, we hypothesized and found high temporal stability of teachers' judgment accuracy. First, regarding student achievement, judgment accuracy did not change significantly between the two measurement points. Our results are in line with prior studies although we tried to avoid self-fulfilling prophecy effects (Hinnant et al., 2009; Lorenz & Artelt, 2009; Oerke et al., 2016; Paleczek et al., 2017). These results indicate that teacher

judgment accuracy as a kind of personal competence is comparatively stable, no matter whether longer or shorter periods are under consideration.

As mentioned earlier, within the four-week test interval, students should not get aware of the expectations that teachers hold for them. The high stability for all kinds of student characteristics confirmed this point of view. When taking a closer look at the changes on the teacher side, however, we found that teachers attempted to modify their assessments and made significant modifications, e.g., to better predict student achievement. Consequently, the value of the differentiation component decreased significantly. Yet, no superior accuracy was found for the rank and level component at the second time of measurement. Taken together, the results document that teachers are trying to abandon perceptual biases and make a conscious effort to reassess students' actual performance.

Second, regarding students' motivation and emotion, teachers' judgment accuracy was consistent over the period of four weeks. These results could supplement the statement that teachers' judgment competence is a rather stable ability in the domains of motivation and emotion (Lorenz & Artelt, 2009). Moreover, teachers' capabilities to accurately judge their students varied considerably across different motivation dimensions. Compared with other assessed attributes, judgment accuracy of students' expectancy of success stayed on a high level, whereas students' test anxiety remained on a low level. To go one step further, teachers tended to adjust their judgments and, in consequence, they described some of the students' motivational and emotional characteristics differently on the second test. In contrast, students' ratings of their motivation and emotion maintained to some degree. The results slightly differ from prior research by Givvin et al. (2001) who found high stability of teachers' judgments about students' motivation over time, but argued that students' self-report motivation tended to be more differentiating and changeable. Teachers' adjustments in the evaluations of students' motivation and emotion may reflect, on the one hand, that teachers have recognized

these situation-specific and less stable variables and are willing to display their awareness. On the other hand, it may also reflect their uncertainty for assessing these hypothetical constructs.

Finally, the structural equation models suggest multiple and strong relations between teachers' judgments and students' characteristics. Teachers' judgments of students' motivation and emotion are robust predictors of their achievement judgments. It should be noticed that teachers were more influenced by their own beliefs about students' motivation than by students' actual motivation (Jussim, 1991; Wijnia, Loyens, Derous, & Schmidt, 2016). Therefore, teachers could predict students' performance quite accurately although students' motivation and emotion were not highly correlated with their actual performance. The high correlations between teachers' judgments of students' achievement and students' motivation and emotion could partly be explained by a halo effect stating the tendency to form consistent impressions of others (Fiske & Neuberg, 1990). Teachers think of a student in general and color their judgments of specific dimensions by this general feeling.

4.4.3 Limitations and Further Directions

Several limitations of the study should also be noted for better interpreting the results. First, teacher expectations and student characteristics were both measured twice within a comparatively short period of time to keep out teacher expectancy effects. Further studies with short time intervals but more measurement points are expected to better describe the development of teacher judgment accuracy. Another limitation of the study is due to the high experience of teachers in the sample. Although very little reliable information is available about the influence of teaching experience on teachers' ability to accurately judge students (Südkamp et al., 2012), it is believed that certain teacher characteristics make it possible to judge students with stable accuracy (Wijnia et al., 2016). The long years of teaching and contact with students could have potentially influenced the outcomes of this study. Thus, it would be interesting to examine fresh teachers' stability of judgment accuracy in future

research. Further, whereas the number of students was sufficiently large, we have to admit that the small numbers of teachers and school limit the generalizability of the findings.

The present study provided several strengths and found evidence for the temporal stability of teacher judgment accuracy on both student achievement and motivation. Some research groups considered different opportunities to enhance teachers' judgment accuracy and have made positive progress. For example, Zhu and Urhahne (2018) reported significant enhancement of teacher judgment accuracy of student achievement with a five-week interval period by use of learner response systems in the classroom. Moreover, Thiede et al. (2018) found effects of professional development programs on the accuracy of teachers' judgments. However, whether teacher judgment accuracy stays at a high level after the intervention remains questionable. Our findings suggest that although positive effects remained during the investigation time frame, the temporal stability of judgment accuracy should be taken into consideration in a future intervention study. There is also considerable evidence suggesting that students' motivation and emotion do not change so easily as usually assumed due to external influences (Bråten & Olaussen, 2005; Givvin et al., 2001; Zhu & Urhahne, 2014). Thus, mainly focusing on students' current moods or what is going on in class at a particular time might not be a good strategy for precise judgments.

4.4.4 Conclusions

In conclusion, teacher judgment accuracy is moderated by the dimension under focus. The best teacher ratings were found for students' academic achievement followed by students' motivation. It seems to be most difficult to properly rate students' emotional traits. Above all, the findings of the study confirm the temporal stability of teacher judgment accuracy for both cognitive and motivational-affective student characteristics. The results also indicate that teachers are aware of students' development and try to adapt. However, they should learn more about valid indicators of students' motivation and emotion if the same degree of judgment accuracy should be reached as in the cognitive domain. To gain more insight into

teachers' judgment process, qualitative studies might be helpful that ask of teachers' strategies for assessing students' affective traits.

CHAPTER 5: GENERAL DISCUSSION

5.1 Summary of Findings

5.1.1 Study 1: Examining Teachers' Strategies to Judge Student Achievement from a Cue Utilization Perspective

In the first study, teachers' strategies to judge student achievement were examined. Social Judgment Theory's lens model was used to interpret teachers' judgment process. According to the lens model, there are mainly three paths to determine an accurate judgment (i.e., cue validity, cue utilization, and the final judgment based on student cues). The research focus was teachers' utilization of different types of student cues.

A semi-structured questionnaire with seven categories and accompanying items was developed according to 34 meta-analyses and previous interviews with 16 primary school teachers (Zhu, 2014). To examine if and to what extent teachers utilize these seven types of student information sources, 260 teachers from seven Chinese primary schools were invited to point out what information in each category they rely on and make an order of the most important three items.

As expected, teachers based their achievement judgments on seven information sources: (a) students' abilities and attitudes, (b) behavior during class, (c) tests, (d) homework, (e) behavior after class, (f) demographics, and (g) other social interactions. The most frequently reported student cues from the seven information sources were students' general intelligence, students' interest, and students' learning strategies (source a); students' engagement during class (concentrates well) (source b); finishing homework punctually, independently and correctly (source d); students' last test performance (source c); questioning and talking with teachers (source e); conversation with parents and colleagues (source g); and parents' educational level (source f).

Consistent with previous research, the results also showed teachers' positive attitudes toward their judgment strategies (Praetorius, Berner, Zeinz, Scheunpflug, & Dresel, 2013). On each of the information sources, teachers were found to rank their most often used student

cues as the most important in the judgment process. It shows that teachers believe that their judgments about student achievement are soundly based.

It was further confirmed that most student cues utilized by teachers are of high validity. For example, as the most important cues for teachers, students' general intelligence and engagement were found to be significantly correlated with students' academic achievement (Lei, Cui, & Zhou, 2018; Roth et al., 2015). From the perspective of Social Judgment Theory, when teachers base their judgments on these student cues, they could predict student achievement precisely. On the other side, some student cues (e.g., students' homework) might be overestimated in the judgment process, which could be a clue to explain inaccurate judgments.

5.1.2 Study 2: The Use of Learner Response Systems in the Classroom Enhances Teachers' Judgment Accuracy

In the second study, it was attempted to enhance teachers' judgment accuracy of student achievement with the use of learner response systems (clickers) in the classroom. As shown in Study 1, teachers' judgment will be more accurate when they utilize valid information about student achievement. The technology of clickers enables teachers to collect and present students' learning outcomes efficiently and it was expected to provide teachers with more detailed information about each individual student. Moreover, student achievement was assumed to be facilitated by the use of clickers.

Twenty German school classes with 459 sixth-grade students and their mathematics teachers were divided into three groups for a quasi-experimental pre-post-test intervention study over five weeks. The clicker classes were equipped with learner response systems and utilized this technology regularly for five weeks. Teachers of the diary group worked on a diary book collecting standardized information about mathematics lessons. Both students and teachers of the control group conducted regular learning and instruction during the period of intervention. Students' mathematics achievement on fractional arithmetic of all three groups

was measured by a standardized mathematics test (DEMAT 6+) in both pre- and post-test. Teachers of all groups were asked to make predictions about students' test performance at both times of measurement.

The findings are in line with the conclusions of prior studies that clickers could positively influence students' learning outcomes (Anderson, Healy, Kole, & Bourne, 2013; Campbell & Monk, 2015; Mayer et al., 2009). The results revealed significantly higher learning gains for students in the clicker classes. The immediate feedback about learning tasks provided to students as well as teachers' adjustment in terms of their instruction and expectations would have contributed to the distinguishable improvement of student achievement in the clicker classes.

The most important finding of this study was the enhancement of judgment accuracy through the use of learner response systems. Teachers of the clicker group were more accurate on the rank component, level component, and the global deviation measure. It could be interpreted that they got benefit from the frequent information about students' current learning outcomes via clickers compared with teachers of the diary and control group. Furthermore, it is noteworthy that the rank component for teachers in the clicker classes increased from $r = .53$ to $r = .90$, indicating an extraordinary higher value than on average reported (Hoge & Coladarci, 1989; Südkamp, Kaiser, & Möller, 2012). This finding constitutes a new approach to empirically improving teacher judgment accuracy. It is also shown that the efforts to try new resources like clickers in class are of great worth.

However, it is unknown yet whether teachers could make judgments with the same accuracy after the withdrawal of clickers. A longitudinal study about the development of teacher judgment accuracy is needed. Related research questions are addressed in Study 3.

5.1.3 Study 3: Temporal Stability of Teachers' Judgment Accuracy of Students' Motivation, Emotion and Achievement

The temporal stability of teacher judgment accuracy, which was seldom explored in previous studies, was considered in the third study. The accuracy of teacher judgment in terms of motivational, emotional, and cognitive student characteristics was examined at two different points of time. It was analyzed how well teachers could judge their students on the three dimensions and whether their judgment accuracy stays on the same level over a four-week time interval. Moreover, structural equation models were applied to examine the interplay of teacher's judgments and students' characteristics.

Data was collected from 326 sixth-graders and their five mathematics teachers of a Chinese elementary school. Students filled in a standardized mathematics test and questionnaire items and scales on motivation and emotion. The mathematics teachers rated each individual student's corresponding test performance and motivational-affective characteristics. The same procedure was conducted after four weeks.

Consistent with earlier study results, teachers were less able to judge students' academic motivation than students' academic performance (Dicke, Lüdtke, Trautwein, Nagy, & Nagy, 2012; Spinath, 2005; Zhu & Urhahne, 2014). They had especially difficulties in predicting students' academic emotions. These results could be partly explained from the perspective of information utilization. There is more available information reflecting students' academic achievement (e.g., students' test records, homework, and exercises) for teachers, whereas information about students' motivation and emotion is more difficult to detect and interpret. Hence, teachers could hardly use cues with high validity to make fair judgments of motivational-affective traits.

Furthermore, teachers' judgment accuracy was found to have high temporal stability for both cognitive and motivational-affective student characteristics. Results showed that teachers tried to make some adjustments of their assessments about students' achievement, self-concept, expectancy of success, and test anxiety at the second time of measurement. However, the judgment accuracy as a whole did not change significantly. This result indicates

the same necessity to conduct some intervention studies to positively influence teachers' judgments about students' motivation and emotion.

5.2 Implications

Research on teachers' judgment has made considerable progress (Shavelson, 1983; Südkamp et al., 2012; Südkamp, Praetorius, & Spinath, 2018). The significance of teacher judgment and the variability in judgment accuracy warrant deeper investigation. The series of studies sought to respond to a new set of important questions related to the processes and features of teacher judgment. In particular, the presented three studies have asked and answered how teachers' judgment is generated from different types of student cues, what are the possible ways to improve teacher judgment accuracy, and whether teachers' judgment accuracy could remain stable over time. The findings have implications for educational practice as well as for future research directions.

5.2.1 Implications for Practice

As one of the most important judgment strategies, it is essential to raise teachers' awareness regarding selection and utilization of student information. When rating students' academic achievement, teachers are reported to utilize student cues from seven information sources, and a part of them was found to be invalid for their judgments. If they rely too heavily on this information with low validity, bias could occur and lead to inaccurate teacher judgment (Bressoux & Pansu, 2016). For example, to make precise predictions of student achievement, teachers are suggested to take students' general intelligence, interest, and engagement into consideration.

The results also indicated that teachers were aware of some stereotype sources and attempted to refrain from them in the judgment process. Teachers' gender stereotypes regarding reading and mathematics, for example, are one of the most often investigated factors in prior studies to examine the influence on teachers' judgments (Baudson, Fischbach, & Preckel, 2016; Holder & Kessels, 2017; Paleczek, Seifert, & Gasteiger-Klicpera, 2017). In

the current study, teachers indicated that students' physical characteristics were not strongly associated with their judgments. To check whether these self-reported claims are true for teachers' actual practice, testing on social desirability bias would be a valuable next step.

Gaining knowledge about students' specific characteristics could foster teachers' judgment accuracy. The use of learner response systems has inspired efforts of improving teachers' judgment accuracy as well as students' academic achievement. As a consequence, teachers are recommended to collect frequent information about students' current performance. More information about students may help them to adjust their expectations and instruction and to provide students with corrective feedback about their learning progress (Anderson et al., 2013; Hattie & Timperley, 2007). In addition, the results of Study 2 provide support for the use of proper technical tools in the classroom (Barnett, 2006). Although it may present some challenges in the beginning, the study shows the benefits of persisting.

Finally, the findings suggest teachers to expand their focus from students' academic outcomes to motivational aspects. It is worth noting that even there are cultural differences that Chinese teachers might place a higher value on students' academic achievement (Gao & Watkins, 2002), both German and Chinese teachers were found to have more difficulties to assess students' motivation and emotions (Urhahne et al., 2010; Zhu & Urhahne, 2014). Therefore it is not enough to just understand the results on the basis of cultural specifics (Zhu & Urhahne, 2014). Teachers should recognize the importance of judgment accuracy regarding students' motivation and emotions, develop effective judgment strategies for detecting specific causes undermining learning, and interpret students' motivation more accurately (Givvin, Stipek, Salmon, & MacGyvers, 2001; Zhu & Urhahne, 2014). Future studies could also assist in providing further support for teachers' rating errors.

5.2.2 Implications for Future Research

While the lens model has been applied to study the path of cue utilization in teachers' judgment process, it could also be used to address other hypotheses for further empirical

testing. For example, research could be extended to measure the three judgment paths with the same experimental sample in an integrated manner. It should examine not only (a) the accuracy of teacher judgment, but also (b) the information sources that teachers use to arrive at accurate judgments, and (c) the validity of student information that teachers incorporate in their judgment processes. Research questions could include the following:

1. How valid are different student information sources for predicting student achievement?

2. What student information sources do teachers rely on to judge student achievement?

3. How accurate are teachers in judging student achievement? Can differences in teacher judgment accuracy be explained by the validity and utilization of different student information sources?

4. Does the use of different student information sources mediate the relationship between teacher judgment of student achievement and students' actual achievement?

As discussed above, further work is necessary to examine teachers' strategies for assessing students' affective traits. The investigation of valid indicators of students' motivation and emotion is scarce, not to mention the studies that aim to provide aids for teachers' judgment accuracy of motivational and emotional variables. Moreover, the qualitative approach could be an alternative to gain insight in the overall formation of teacher's judgment.

As a final point, it is advisable for the future intervention study to take the temporal stability into consideration as teacher judgment accuracy on both student achievement and motivation was highly stable over time. The questions of how teacher judgment accuracy develops and whether it stays at a high level after the experimental intervention or training program are expected to be answered.

REFERENCES

Allison, P.D. (2002). *Missing Data*. Thousand Oaks: Sage.

Alvidrez, J., & Weinstein, R. S. (1999). Early teacher perceptions and later student academic achievement. *Journal of Educational Psychology, 91*, 731–746. doi: 10.1037/0022-0663.91.4.731

Anders, Y., Kunter, M., Brunner, M., Krauss, S., & Baumert, J. (2010). Diagnostische Fähigkeiten von Mathematiklehrkräften und ihre Auswirkungen auf die Leistungen ihrer Schülerinnen und Schüler [Mathematics teachers' diagnostic skills and their impact on students' achievement]. *Psychologie in Erziehung und Unterricht, 57*, 175–193. doi: 10.2378/peu2010.art13d

Anderson, L. S., Healy, A. F., Kole, J. A., & Bourne, L. E., Jr. (2013). The clicker technique: Cultivating efficient teaching and successful learning. *Applied Cognitive Psychology, 27*, 222–234. doi: 10.1002/acp.2899

Arbuckle, J. L. (2017). *IBM SPSS Amos 25. User's Guide*. Chicago: SPSS.

Babad, E. (1993). Teachers' differential behavior. *Educational Psychology Review, 5*, 347–376. doi: 10.1007/BF01320223

Barnett, J. (2006). Implementation of personal response units in very large lecture classes: Student perceptions. *Australasian Journal of Educational Technology, 22*, 474–494. doi: 10.14742/ajet.1281

Bates, C., & Nettelbeck, T. (2001). Primary school teachers' judgements of reading achievement. *Educational Psychology, 21*, 177–187. doi: 10.1080/01443410020043878

Baudson, T. G., Fischbach, A., & Preckel, F. (2016). Teacher judgments as measures of children's cognitive ability: A multilevel analysis. *Learning and Individual Differences, 52*, 148–156. doi: 10.1016/j.lindif.2014.06.001

Bennett, R. E., Gottesman, R. L., Rock, D. A., & Cerullo, F. (1993). Influence of behavior perceptions and gender on teachers' judgments of students' academic skill. *Journal of Educational Psychology, 85*(2), 347–356. doi: 10.1037/0022-0663.85.2.347

Bergin, D. A. (1999). Influences on classroom interest. *Educational Psychologist, 34*, 87–98. doi: 10.1207/s15326985ep3402_2

Bos, W., & Tarnai, C. (1999). Content analysis in empirical social research. *International journal of educational research, 31*, 659–671. doi: 10.1016/S0883-0355(99)00032-4

Bråten, I., & Olaussen, B. S. (2005). Profiling individual differences in student motivation: A longitudinal cluster-analytic study in different academic contexts. *Contemporary Educational Psychology, 30*, 359–396. doi: 10.1016/j.cedpsych.2005.01.003

Bressoux, P., & Pansu, P. (2016). Pupils' self-perceptions: the role of teachers' judgment controlling for big-fish-little-pond effect. *European Journal of Psychology of Education, 31*, 341–357. doi: 10.1007/s10212-015-0264-7

Brophy, J. (1983). Research on the self-fulfilling prophecy and teacher expectations. *Journal of Educational Psychology, 75*, 631–661. doi: 10.1037/0022-0663.75.5.631

Brophy, J., & Good, T. (1970). Teachers' communication of differential expectations for children's classroom performance: Some behavioral data. *Journal of Educational Psychology, 61*, 365–374. doi: 10.1037/h0029908

Bruff, D. (2009). *Teaching with classroom response systems: creating active learning environments*. San Francisco: Jossey-Bass.

Brunswik, E. (1955). Representative design and probabilistic theory in a functional psychology. *Psychological Review, 62*, 193–217. doi: 10.1037/h0047470

Buil, I., Catalán, S., & Martínez, E. (2016). Do clickers enhance learning? A control-value theory approach. *Computers & Education, 103*, 170–182. doi: 10.1016/j.compedu.2016.10.009

Campbell, C., & Monk, S. (2015). Introducing a learner response system to pre-service education students: Increasing student engagement. *Active Learning in Higher Education, 16*, 25–36. doi: 10.1177/1469787414558981

Castro, M., Expósito-Casas, E., López-Martín, E., Lizasoain, L., Navarro-Asencio, E., & Gaviria, J. L. (2015). Parental involvement on student academic achievement: A meta-analysis. *Educational Research Review, 14*, 33–46. doi: 10.1016/j.edurev.2015.01.002

Cerasoli, C. P., Nicklin, J. M., & Ford, M. T. (2014). Intrinsic motivation and extrinsic incentives jointly predict performance: A 40-year meta-analysis. *Psychological Bulletin, 140*, 980–1008. doi: 10.1037/a0035661

Chamorro-Premuzic, T., & Furnham, A. (2003). Personality predicts academic performance: Evidence from two longitudinal university samples. *Journal of Research in Personality, 37*, 319–338. doi: 10.1016/S0092-6566(02)00578-0

Chen, F. F. (2007). Sensitivity of goodness of fit indices to lack of measurement invariance. *Structural Equation Modeling, 14*, 464–504. doi: 10.1080/1070551070130834

Cheung, G. W., & Rensvold, R. B. (2002). Evaluating goodness-of-fit indexes for testing measurement invariance. *Structural Equation Modeling, 9*, 233–255. doi: 10.1207/S15328007SEM0902_5

Chien, Y., Chang, Y., & Chang, C. (2016). Do we click in the right way? A meta-analytic review of clicker-integrated instruction. *Educational Research Review, 17*, 1–18. doi: 10.1016/j.edurev.2015.10.003

Chien, Y. T., Lee, Y. H., Li, T. Y., & Chang, C. Y. (2015). Examining the effects of displaying clicker voting results on high school students' voting behaviors, discussion processes, and learning outcomes. *Eurasia Journal of Mathematics, Science & Technology Education*, *11*, 1089–1104. doi: 10.12973/eurasia.2015.1414a

Cohen, J. (1992). A power primer. *Psychological Bulletin, 112*, 155–159. doi:10.1037/0033-2909.112.1.155

Connor, L. T., Dunlosky, J., & Hertzog, C. (1997). Age-related differences in absolute but not relative metamemory accuracy. *Psychology and Aging, 12*, 50–71. doi: 10.1037/0882-7974.12.1.50

Cooksey, R. W., Freebody, P., & Davidson, G. R. (1986). Teachers' predictions of children's early reading achievement: An application of social judgment theory. *American Educational Research Journal, 23*, 41–64. doi: 10.3102/00028312023001041

Cooper, H., Robinson, J. C., & Patall, E. A. (2006). Does homework improve academic achievement? A synthesis of research, 1987-2003. *Review of Educational Research, 76*, 1–62. doi: 10.3102/00346543076001001

Cronbach, L. J. (1955). Processes affecting scores on "understanding of others" and "assumed similarity." *Psychological Bulletin, 52*, 177–193. doi: 10.1037/h0044919

Dalbert, C., Schneidewind, U., & Saalbach, A. (2007). Justice judgments concerning grading in school. *Contemporary Educational Psychology, 32*, 420–433. doi: 10.1016/j.cedpsych.2006.05.003

Dent, A. L., & Koenka, A. C. (2016). The relation between self-regulated learning and academic achievement across childhood and adolescence: A meta-analysis. *Educational Psychology Review, 28*, 425–474. doi: 10.1007/s10648-015-9320-8

Dicke, A.-L., Lüdtke, O., Trautwein, U., Nagy, G., & Nagy, N. (2012). Judging students' achievement goal orientations: Are teacher ratings accurate? *Learning and Individual Differences, 22*, 844–849. doi: 10.1016/j.lindif.2012.04.004

Dietrichson, J., Bøg, M., Filges, T., & Klint Jørgensen, A. M. (2017). Academic interventions for elementary and middle school students with low socioeconomic status: A systematic review and meta-analysis. *Review of Educational Research, 87*, 243–282. doi: 10.3102/0034654316687036

Doherty, J., & Conolly, M. (1985). How accurately can primary school teachers predict the scores of their pupils in standardised tests of attainment? A study of some non-

cognitive factors that influence specific judgements. *Educational Studies, 11*, 41–60. doi: 10.1080/0305569850110105

Duncan, G. J., Dowsett, C. J., Claessens, A., Magnuson, K., Huston, A. C., Klebanov, P., . . . Japel, C. (2007). School readiness and later achievement. *Developmental Psychology, 43*, 1428–1446. doi: 10.1037/0012-1649.43.6.1428

Dusek, J. B., & Joseph, G. (1983). The bases of teacher expectancies: A meta-analysis. *Journal of Educational Psychology, 75*, 327–346. doi: 10.1037/0022-0663.75.3.327

Faber, J. M., Luyten, H., & Visscher, A. J. (2017). The effects of a digital formative assessment tool on mathematics achievement and student motivation: results of a randomized experiment. *Computer & Education, 106*, 83–96. doi: 10.1016/j.compedu.2016.12.001

Fan, H., Xu, J., Cai, Z., He, J., & Fan, X. (2017). Homework and students' achievement in math and science: A 30-year meta-analysis, 1986-2015. *Educational Research Review, 20*, 35–54. doi: 10.1016/j.edurev.2016.11.003

Fan, X., & Chen, M. (2001). Parental involvement and students' academic achievement: A meta-analysis. *Educational Psychology Review, 13*, 1–22. doi: 10.1023/A:1009048817385

Fiske, S. T., & Neuberg, S. L. (1990). A continuum of impression formation, from category-based to individuating processes: Influences of information and motivation on attention and interpretation. *Advances in Experimental Social Psychology, 23*, 1–74. doi: 10.1016/S0065-2601(08)60317-2

Förster, N., & Souvignier, E. (2017). Förderung diagnostischer Kompetenz durch Bereitstellung formativer Diagnostik [Promoting diagnostic competence through the provision of formative diagnostics]. In A. Südkamp & A.-K. Praetorius (Eds.), *Diagnostische Kompetenz von Lehrkräften. Theoretische und methodische Weiterentwicklungen* (pp. 231–239). Münster: Waxmann.

Franke, R. H., & Kaul, J. D. (1978) The Hawthorne experiments: First statistical interpretation. *American Sociological Review, 43,* 623–643. doi: 10.2307/2094540

Friedrich, A. Flunger, B., Nagengast, B., Jonkmann, K., & Trautwein, U. (2015). Pygmalion effects in the classroom: Teacher expectancy effects on students' math achievement. *Contemporary Educational Psychology, 41*, 1–12. doi: 10.1016/j.cedpsych.2014.10.006

Frieze, I. H., & Weiner, B. (1971). Cue utilization and attributional judgments for success and failure. *Journal of Personality, 39,* 591–605. doi: 10.1111/j.1467-6494.1971.tb00065.x

Funder, D. C. (2012). Accurate personality judgment. *Current Directions in Psychological Science, 21*, 177–182. doi: 10.1177/0963721412445309

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. doi: 10.1016/j.learninstruc.2016.06.008

Gao, L., & Watkins, D. A. (2002). Conceptions of teaching held by school science teachers in P.R. China: Identification and cross-cultural comparisons. *Science Education, 24,* 61–79. doi: 10.1080/09500690110066926

Garson, G. D. (Ed.). (2013). *Hierarchical linear modeling. Guide and applications*. Thousand Oaks, CA: Sage.

Givvin, K. B., Stipek, D. J., Salmon, J. M., MacGyvers, V. L. (2001). In the eyes of the beholder: students' and teachers' judgments of students' motivation. *Teaching and Teacher Education, 17,* 321–331. doi: 10.1016/S0742-051X(00)00060-3

Glogger-Frey, I., Herppich, S., & Seidel, T. (2018). Linking teachers' professional knowledge and teachers' actions: Judgment processes, judgments and training. *Teaching and Teacher Education, 76*, 176–180. doi: 10.1016/j.tate.2018.08.005

Glogger-Frey, I., & Renkl, A. (2017). Diagnostische Kompetenz fördern – Vorwissen aufgreifende Methoden in Kombination mit beispielbasiertem Kurztraining [Promoting diagnostic competence – Knowledge-based methods combined with example-based short training]. In A. Südkamp & A.-K. Praetorius (Eds.), *Diagnostische Kompetenz von Lehrkräften. Theoretische und methodische Weiterentwicklungen* (pp. 217-222). Münster: Waxmann.

Götz, L., Lingel, K., & Schneider, W. (2013). *DEMAT 6+. Deutscher Mathematiktest für sechste Klassen* [German mathematics test for the sixth grade]. Göttingen: Hogrefe.

Good, T. L. (1987). Two decades of research on teacher expectations: Findings and future directions. *Journal of Teacher Education, 38,* 32–47. doi: 10.1177/002248718703800406

Gray, S. A., Dueck, K., Rogers, M., & Tannock, R. (2017). Qualitative review synthesis: The relationship between inattention and academic achievement. *Educational Research, 59,* 17–35. doi: 10.1080/00131881.2016.1274235

Haigh, M., & Ell, F. (2014). Consensus and dissensus in mentor teachers' judgments of readiness to teach. *Teaching and Teacher Education, 40,* 10–21. doi: 10.1016/j.tate.2014.01.001

Haigh, M., Ell, F., & Mackisack, V. (2013). Judging teacher candidates' readiness to teach. *Teaching and Teacher Education, 34,* 1–11. doi: 10.1016/j.tate.2013.03.002

Hammond, K., Rohrbaugh, J., Mumpower, J., & Adelman, L. (1977). Social Judgment Theory: Applications in policy formation. In M. Kaplan, & S. Schwartz (Eds.), *Human judgment and decision processes in applied settings* (pp. 1–29). New York: Academic Press.

Hattie, J. (2012*). Visible learning for teachers. Maximizing impact on learning*. New York: Routledge.

Hattie, J., & Timperley, H. (2007). The power of feedback. *Review of Educational Research, 77*, 81–112. doi: 10.3102/003465430298487

He, Q., Valcke, M., & Aelterman, A. (2012). A qualitative study of in-service teacher evaluation beliefs. *Journal of Educational Sciences & Psychology, 2*, 1–14. doi: 10.1016/j.sbspro.2012.12.035

Heald, J. E. (1991). Social judgment theory: Applications to educational decision making. *Educational Administration Quarterly, 27*, 343–357. doi: 10.1177/0013161X91027003006

Hecht, S. A., & Greenfield, D. B. (2002). Explaining the predictive accuracy of teacher judgments of their students' reading achievement: The role of gender, classroom behavior, and emergent literacy skills in a longitudinal sample of children exposed to poverty. *Reading and Writing, 15*, 789–809. doi: 10.1023/A:1020985701556

Helm, F., Müller-Kalthoff, H., Mukowski, R., & Möller, J. (2018). Teacher judgment accuracy regarding students' self-concepts: Affected by social and dimensional comparisons? *Learning and Instruction, 55*, 1–12. doi: 10.1016/j.learninstruc.2018.02.002

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

Hinnant, J. B., O'Brien, M., & Ghazarian, S. R. (2009). The longitudinal relations of teacher expectations to achievement in the early school years. *Journal of Educational Psychology, 101*, 662–670. doi: 10.1037/a0014306

Hoge, R. D., & Coladarci, T. (1989). Teacher-based judgment of academic achievement: A review of literature. *Review of Educational Research, 59*, 297–313. doi: 10.3102/00346543059003297

Hoge, R. D., & Cudmore, L. (1986). The use of teacher-judgment measures in the identification of gifted pupils. *Teaching & Teacher Education, 2*, 181–196. doi: 10.1016/0742-051X(86)90016-8

Holder, K., & Kessels, U. (2017). Gender and ethnic stereotypes in student teachers' judgments: A new look from a shifting standards perspective. *Social Psychology of Education, 20*, 471–490. doi: 10.1007/s11218-017-9384-z

Hu, L. T., & Bentler, P. M. (1999). Cutoff criteria for fit indexes in covariance structure analysis: Conventional criteria versus new alternatives. *Structural Equation Modeling, 6*, 1–55. doi: 10.1080/10705519909540118

Huang, C. (2011). Self-concept and academic achievement: A meta-analysis of longitudinal relations. *Journal of School Psychology, 49*, 505–528. doi: 10.1016/j.jsp.2011.07.001

Huang, C. (2012). Discriminant and criterion-related validity of achievement goals in predicting academic achievement: A meta-analysis. *Journal of Educational Psychology, 104*, 48–73. doi: 10.1037/a0026223

Huber, S. A., & Seidel, T. (2018). Comparing teacher and student perspectives on the interplay of cognitive and motivational-affective student characteristics. *PLoS ONE, 13*(8), e0200609. doi: 10.1371/journal.pone.02006009

Hulleman, C. S., Schrager, S. M., Bodmann, S. M., & Harackiewicz, J. M. (2010). A meta-analytic review of achievement goal measures: Different labels for the same constructs or different constructs with similar labels? *Psychological Bulletin, 136*, 422–449. doi: 10.1037/a0018947

Hunsu, N. J., Adesope, O., & Bayly, D. J. (2016). A meta-analysis of the effects of audience response systems (clicker-based technologies) on cognition and affect. *Computers & Education, 94*, 102–119. doi: 10.1016/j.compedu.2015.11.013

Jamil, F. M., Larsen, R. A., & Hamre, B. K. (2018). Exploring longitudinal changes in teacher expectancy effects on children's mathematics achievement. *Journal of Research in Mathematics Education, 4*9, 57–90. doi: 10.5951/jresematheduc.49.1.0057

Jurik, V., Gröschner, A., & Seidel, T. (2013). How student characteristics affect girls' and boys' verbal engagement in physics instruction. *Learning and Instruction, 23*, 33–42. doi: 10.1016/j.learninstruc.2012.09.002

Jussim, L. (1989). Teacher expectations: Self-fulfilling prophecies, perceptual biases, and accuracy. *Journal of Personality and Social Psychology, 57*, 469–480. doi: 10.1037/0022-3514.57.3.469

Jussim, L. (1991). Social perception and social reality: A reflection-construction model. *Psychological Review, 98*, 54–73. doi: 10.1037/0033-295X.98.1.54

Jussim, L., & Eccles, J. (1995). Naturalistic studies of interpersonal expectancies. *Review of Personality and Social Psychology, 15*, 74–108.

Jussim, L., & Harber, K. D. (2005). Teacher expectations and self-fulfilling prophecies: Knowns and unknowns, resolved and unresolved controversies. *Personality and Social Psychology Review, 9*, 131–155. doi: 10.1207/s15327957pspr0902_3

Kaiser, J., Möller, J., Helm, F., & Kunter, M. (2015). Das Schülerinventar: Welche Schülermerkmale die Leistungsurteile von Lehrkräften beeinflussen [The student inventory: how student characteristics bias teacher judgments]. *Zeitschrift für Erziehungswissenschaft, 18*, 279–302. doi: 10.1007/s11618-015-0619-5

Kaiser, J., Retelsdorf, J., Südkamp, A., & Möller, J. (2013). Achievement and engagement: How student characteristics influence teacher judgements. *Learning and Instruction, 28*, 73–84. doi: 10.1016/j.learninstruc.2013.06.001

Kaiser, J., Südkamp, A., & Möller, J. (2017). The effects of student characteristics on teachers' judgment accuracy: Disentangling ethnicity, minority status, and achievement. *Journal of Educational Psychology, 109*, 871–888. doi: 10.1037/edu0000156

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

Karing, C., Dörfler, T., & Artelt, C. (2015). How accurate are teacher and parent judgements of lower secondary school children's test anxiety? *Educational Psychology, 35,* 909–925. doi: 10.1080/01443410.2013.814200

Kelley, H. H. (1967). Attribution theory in social psychology. *Nebraska Symposium on Motivation, 15,* 192–238.

Keough, S. M. (2012). Clickers in the classroom: A review and a replication. *Journal of Management Education, 36,* 822–847. doi: 10.1177/1052562912454808

Kim, K. R., & Seo, E. H. (2015). The relationship between procrastination and academic performance: A meta-analysis. *Personality and Individual Differences, 82,* 26–33. doi: 10.1016/j.paid.2015.02.038

Kim, S., & Fong, V. L. (2013). How parents help children with homework in China: Narratives across the life span. *Asia Pacific Education Review, 14,* 581–592. doi: 10.1007/s12564-013-9284-7

Kishor, N. (1989). Information utilization in performance rating: Interactive effects of purpose and cue dimensionality. *Journal of Personnel Evaluation in education, 2,* 177–191. doi: 10.1007/BF00127179

Kishor, N. (1994). Teachers' judgements of students' performance: use of consensus, consistency and distinctiveness information. *Educational Psychology, 14,* 233–247. doi: 10.1080/0144341940140207

Klug, J., Bruder, S., & Schmitz, B. (2016). Which variables predict teachers diagnostic competence when diagnosing students' learning behavior at different stages of a

teacher's career? *Teachers and Teaching, 22*, 461–484. doi: 10.1080/13540602.2015.1082729

Klug, J., Gerich, M., Bruder, S., & Schmitz, B. (2012). Ein Tagebuch für Hauptschullehrkräfte zur Unterstützung der Reflektionsprozesse beim Diagnostizieren [A teachers' diary to support reflection processes during diagnosing]. *Empirische Pädagogik, 26*, 292–311.

Klug, J., Gerich, M., & Schmitz, B. (2016). Can teachers' diagnostic competence be fostered through training and the use of a diary? *Journal for Educational Research Online, 8*(3), 184–206.

Kraft, M. A., & Rogers, T. (2015). The underutilized potential of teacher-to-parent communication: Evidence from a field experiment. *Economics of Education Review, 47*, 49–63. doi: 10.1016/j.econedurev.2015.04.001

Lantz, M. E. (2010). The use of 'clickers' in the classroom: Teaching innovation or merely an amusing novelty?. *Computers in Human Behavior, 26*, 556–561. doi: 10.1016/j.chb.2010.02.014

Lantz, M. E., & Stawiski, A. (2014). Effectiveness of clickers: Effect of feedback and the timing of questions on learning. *Computers in Human Behavior, 31*, 280–286. doi: 10.1016/j.chb.2013.10.009

Lauer, P. A., Akiba, M., Wilkerson, S. B., Apthorp, H. S., Snow, D., & Martin-Glenn, M. L. (2006). Out-of-school-time programs: A meta-analysis of effects for at-risk students. *Review of Educational Research, 76*, 275–313. doi: 10.3102/00346543076002275

Lee, J., & Stankov, L. (2018). Non-cognitive predictors of academic achievement: Evidence from TIMSS and PISA. *Learning and Individual Differences, 65*, 50–64. doi: 10.1016/j.lindif.2018.05.009

Lei, H., & Cui, Y. (2016). Effects of academic emotions on achievement among mainland Chinese students: A meta-analysis. *Social Behavior and Personality: an international journal, 44*, 1541–1553. doi: 10.2224/sbp.2016.44.9.1541

Lei, H., Cui, Y., & Zhou, W. (2018). Relationships between student engagement and academic achievement: A meta-analysis. *Social Behavior and Personality: an international journal, 46*, 517–528. doi: 10.2224/sbp.7054

Lietz, P. (2006). A meta-analysis of gender differences in reading achievement at the secondary school level. *Studies in Educational Evaluation, 32*, 317–344. doi: 10.1016/j.stueduc.2006.10.002

Lindberg, S. M., Hyde, J. S., Petersen, J. L., & Linn, M. C. (2010). New trends in gender and mathematics performance: A meta-analysis. *Psychological Bulletin, 136*, 1123–1135. doi: 10.1037/a0021276

Lissmann, U. (2010). *Leistungsmessung und Leistungsbeurteilung – eine Einführung* [Performance measurement and performance evaluation – An introduction]. Landau: Verlag Empirische Pädagogik.

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

Ma, X. (1999). A meta-analysis of the relationship between anxiety toward mathematics and achievement in mathematics. *Journal for Research in Mathematics Education*, 520–540. doi: 10.2307/749772

Machts, N., Kaiser, J., Schmidt, F. T., & Moeller, J. (2016). Accuracy of teachers' judgments of students' cognitive abilities: A meta-analysis. *Educational Research Review, 19*, 85–103. doi: 10.1016/j.edurev.2016.06.003

Madelaine, A., & Wheldall, K. (2005). Identifying low-progress readers: Comparing teacher judgment with a curriculum-based measurement procedure. *International Journal of Disability Development and Education, 52,* 3–42. doi: 10.1080/10349120500071886

Malouff, J. M., & Thorsteinsson, E. B. (2016). Bias in grading: A meta-analysis of experimental research findings. *Australian Journal of Education, 60,* 245–256. doi: 10.1177/0004944116664618

Marjoribanks, K. (1987). Ability and attitude correlates of academic achievement: Family-group differences. *Journal of Educational Psychology, 79,* 171–178. doi: 10.1037/0022-0663.79.2.171

Mayer, R. E., Stull, A., Deleeuw, K., Almeroth, K., Bimber, B., Chun, D., et al. (2009). Clickers in college classrooms: Fostering learning with questioning methods in large lecture classes. *Contemporary Educational Psychology, 34,* 51–57. doi: 10.1016/j.cedpsych.2008.04.002

McDonough, K., & Foote, J. A. (2015). The impact of individual and shared clicker use on students' collaborative learning. *Computers & Education, 86,* 236–249. doi: 10.1016/j.compedu.2015.08.009

Meissel, K., Meyer, F., Yao, E. S., & Rubie-Davies, C. M. (2017). Subjectivity of teacher judgments: Exploring student characteristics that influence teacher judgments of student ability. *Teaching and Teacher Education, 65,* 48–60. doi: 10.1016/j.tate.2017.02.021

Möller, J., Pohlmann, B., Köller, O., & Marsh, H. W. (2009). A meta-analytic path analysis of the internal/external frame of reference model of academic achievement and academic self-concept. *Review of Educational Research, 79,* 1129–1167. doi: 10.3102/0034654309337522

Murayama, K., & Elliot, A. J. (2012). The competition-performance relation: A meta-analytic review and test of the opposing processes model of competition and performance. *Psychological Bulletin, 138*, 1035–1070. doi: 10.1037/a0028324

Nestler, S., & Back, M. D. (2013). Applications and extensions of the lens model to understand interpersonal judgments at zero acquaintance. *Current Directions in Psychological Science, 22*, 374–379. doi: 10.1177/0963721413486148

Oerke, B., McElvany, N., Ohle, A., Ullrich, M., & Horz, H. (2016). Verbessert sich die diagnostische Urteilsgenauigkeit von Lehrkräften bei längerem Kontakt mit der Klasse? [Does diagnostic accuracy of teachers improve with longer contact to the class?], *Psychologie in Erziehung und Unterricht, 63*, 34–47. doi: 10.2378/peu2016.art04d

Oudman, S., van de Pol, J., Bakker, A., Moerbeek, M., & van Gog, T. (2018). Effects of different cue types on the accuracy of primary school teachers' judgments of students' mathematical understanding. *Teaching and Teacher Education.* Advance online publication. doi: 10.1016/j.tate.2018.02.007

Paleczek, L. Seifert, S., & Gasteiger-Klicpera, B. (2017). Influences on teachers' judgment accuracy of reading abilities on second and third grade students: A multilevel analysis. *Psychology in the Schools, 54*, 228–245. doi: 10.1002/pits.21993

Pang, I. W., & Watkins, D. (2000). Towards a psychological model of teacher-parent communication in Hong Kong primary schools. *Educational Studies, 26*, 141–163. doi: 10.1080/713664272

Pekrun, R., Frenzel, A. C., Goetz, T., & He, S. (2005). *Achievement Emotions Questionnaire-Mathematics (AEQ-M). Chinese version. User's manual*. Munich: University of Munich.

Pekrun, R., Muis, K. R., Frenzel, A. C., & Goetz, T. (2018). *Emotions at school*. New York: Routledge.

Petscher, Y. (2010). A meta-analysis of the relationship between student attitudes towards reading and achievement in reading. *Journal of Research in Reading*, *33*(4), 335–355. doi: 10.1111/j.1467-9817.2009.01418.x

Pit-ten Cate, I. M., Krolak-Schwerdt, S., & Glock, S. (2016). Accuracy of teachers' tracking decisions: Short- and long-term effects of accountability. *European Journal of Psychology of Education, 31*, 225–243. doi: 10.1007/s10212-015-0259-4

Praetorius, A. K., Berner, V. D., Zeinz, H., Scheunpflug, A., & Dresel, M. (2013). Judgment confidence and judgment accuracy of teachers in judging self-concepts of students. *The Journal of Educational Research, 106*, 64–76. doi: 10.1080/00220671.2012.667010

Praetorius, A. K., Koch, T., Scheunpflug, A., Zeinz, H., & Dresel, M. (2017). Identifying determinants of teachers' judgment (in)accuracy regarding students' school-related motivations using a Bayesian cross-classified multi-level model. *Learning and Instruction, 52*, 148–160. doi: 10.1016/j.learninstruc.2017.06.003

Rammstedt, B., & Rammsayer, T. H. (2002). Die Erfassung von selbsteingeschätzter Intelligenz: Konstruktion, teststatistische Überprüfung und erste Ergebnisse des ISI [Assessment of self-estimated intelligence: construction, statistical testing, and first results of the Inventory of Self-Estimated Intelligence (ISI)]. *Zeitschrift für Differentielle und Diagnostische Psychologie, 23*, 435–446. doi: 10.1024//0170-1789.23.4.435

Rausch, T., Karing, C., Dörfler, T., & Artelt, C. (2016). Personality similarity between teachers and their students influences teacher judgement of student achievement. *Educational Psychology, 36*, 863–878. doi: 10.1080/01443410.2014.998629

Reilly, D., Neumann, D. L., & Andrews, G. (2015). Sex differences in mathematics and science achievement: A meta-analysis of National Assessment of Educational Progress

assessments. *Journal of Educational Psychology, 107*, 645–662. doi: 10.1037/edu0000012

Rich, J. (1975). Effects of children's physical attractiveness on teachers' evaluations. *Journal of Educational Psychology, 67*, 599–609. doi: 10.1037/0022-0663.67.5.599

Richardson, A. M., Dunn, P. K., McDonald, C., & Oprescu, F. (2015). CRiSP: an instrument for assessing student perceptions of classroom response systems. *Journal of Science Education and Technology, 24*, 432–447. doi: 10.1007/s10956-014-9528-2

Richardson, M., Abraham, C., & Bond, R. (2012). Psychological correlates of university students' academic performance: A systematic review and meta-analysis. *Psychological Bulletin, 138*, 353–387. doi: 10.1037/a0026838

Roediger, H. L., & Karpicke, J. (2006). The power of testing memory: Basic research and implications for educational practice. *Perspectives on Psychological Science, 1*, 181–210. doi: 10.1111/j.1745-6916.2006.00012.x

Roorda, D. L., Koomen, H. M., Spilt, J. L., & Oort, F. J. (2011). The influence of affective teacher-student relationships on students' school engagement and achievement: A meta-analytic approach. *Review of Educational Research, 81*, 493–529. doi: 10.3102/0034654311421793

Rosenthal, R. (1991). Teacher expectancy effects: A brief update 25 years after the Pygmalion Experiment. *Journal of Research in Education, 1*, 3–12.

Rosenthal, R., & Jacobson, L. (1968). *Pygmalion in the classroom: Teacher expectations and student intellectual development*. New York: Holt, Rinehart, and Winston.

Roth, B., Becker, N., Romeyke, S., Schäfer, S., Domnick, F., & Spinath, F. M. (2015). Intelligence and school grades: A meta-analysis. *Intelligence, 53*, 118–137. doi：10.1016/j.intell.2015.09.002

Rubie-Davies, C. M. (2010). Teacher expectations and perceptions of student attributes: Is there a relationship? *British Journal of Educational Psychology, 80*, 121–135. doi: 10.1348/000709909X466334

Sacher, W., & Rademacher, S. (2009). *Leistungen entwickeln, überprüfen und beurteilen. Bewährte und neue Wege für die Primar- und Sekundarstufe* [Develop, review, and assess achievement. Proven and new paths for primary and secondary education]. Bad Heilbrunn: Klinkhardt.

Schafer, J. L., & Graham, J. W. (2002). Missing data: Our view of the state of the art. *Psychological Methods, 7*(2), 147–177. doi: 10.1037/1082-989X.7.2.147

Schiefele, U., Krapp, A., & Schreyer, I. (1993). Metaanalyse des Zusammenhangs von Interesse und schulischer Leistung [A metaanalysis of the interaction between scholastic interest and scholastic achievement]. *Zeitschrift für Entwicklungspsychologie und Pädagogische Psychologie*, *25*, 120–148.

Schiefele, U., Krapp, A., & Winteler, A. (1992). Interest as a predictor of academic achievement: A meta-analysis of research. In K. A. Renninger, S. Hidi, & A. Krapp (Eds.), *The role of interest in learning and development* (pp. 183–212). Hillsdale: Erlbaum.

Schmitz, B., & Wiese, B. S. (2006). New perspectives for the evaluation of training sessions in self-regulated learning: Time-series analyses of diary data. *Contemporary Educational Psychology, 31*, 64–96. doi: 10.1016/j.cedpsych.2005.02.002

Shavelson, R. J. (1983). Review of research on teachers' pedagogical judgment, plans, and decisions. *The Elementary School Journal, 83*, 392–413.

Shavelson, R. J., & Stern, P. (1981). Research on teachers' pedagogical thoughts, judgments, decisions, and behavior. *Review of Educational Research, 51*, 455–498. doi: 10.3102/00346543051004455

Shulman, L. S., & Elstein, A. S. (1975). Studies of problem solving, judgment, and decision making: Implications for educational research. *Review of Research in Education, 3*, 3–42. doi: 10.3102/0091732X003001003

Singer, V., & Strasser, K. (2017). The association between arithmetic and reading performance in school: A meta-analytic study. *School Psychology Quarterly, 32*, 435–448. doi: 10.1037/spq0000197

Sirin, S. R. (2005). Socioeconomic status and academic achievement: A meta-analytic review of research. *Review of Educational Research, 75*, 417–453. doi: 10.3102/00346543075003417

Snow, R. E., Corno, L., & Jackson, D. (1996). Individual differences in affective and conative functions. In D. C. Berliner, & R. C. Calfee (Eds.), *Handbook of Educational Psychology* (pp. 243–310). New York: MacMillan.

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*, 85–95. doi: 10.1024/1010-0652.19.12.85

Stang, J., & Urhahne, D. (2016). Stabilität, Bezugsnormorientierung und Auswirkungen der Urteilsgenauigkeit [Stability, reference norm orientation, and effects of judgment accuracy]. *Zeitschrift für Pädagogische Psychologie, 30*, 251–262. doi: 10.1024/1010-0652/a000190

St-Onge, C., Chamberland, M., Lévesque, A., & Varpio, L. (2016). Expectations, observations, and the cognitive processes that bind them: Expert assessment of examinee performance. *Advances in Health Sciences Education, 21*, 627–642. doi: 10.1007/s10459-015-9656-3

Südkamp, A., Kaiser, J., & Möller, J. (2012). Accuracy of teachers' judgments of students' academic achievement: A meta-analysis. *Journal of Educational Psychology, 104*, 743–762. doi: 10.1037/a0027627

Südkamp, A., & Möller, J. (2009). Referenzgruppeneffekte im Simulierten Klassenraum: Direkte und indirekte Einschätzungen von Schülerleistungen [Reference-group-effects in a simulated classroom: Direct and indirect judgments]. Zeitschrift für Pädagogische Psychologie, 23, 161–174. doi: 10.1024/1010-0652.23.34.161

Südkamp, A., Möller, J., & Pohlmann, B. (2008). Der Simulierte Klassenraum: Eine experimentelle Untersuchung zur diagnostischen Kompetenz [The simulated classroom: An experimenal investigation on diagnostic competence]. *Zeitschrift für Pädagogische Psychologie, 22*, 261–276. doi:10.1024/1010-0652.22.34.261

Südkamp, A. & Praetorius, A.-K. (Hrsg.). (2017). *Diagnostische Kompetenz von Lehrkräften*. Münster: Waxmann.

Südkamp, A., Praetorius, A.-K., & Spinath, B. (2018). Teachers' judgment accuracy concerning consistent and inconsistent student profiles. *Teaching and Teacher Education, 76*, 204–213. doi: 10.1016/j.tate.2017.09.016

Talsma, K., Schüz, B., Schwarzer, R., & Norris, K. (2018). I believe, therefore I achieve (and vice versa): A meta-analytic cross-lagged panel analysis of self-efficacy and academic performance. *Learning and Individual Differences, 61*, 136–150. doi: 10.1016/j.lindif.2017.11.015

Thiede, K. W., Brendefur, J. L., Carney, M. B., Champion, J., Turner, L., Stewart, R., & Osguthorpe, R. D. (2018). Improving the accuracy of teachers' judgments of student learning. *Teaching and Teacher Education, 76*, 106–115. doi: 10.1016/j.tate.2018.08.004

Thiede, K. W., Brendefur, J. L., Osguthorpe, R. D., Carney, M. B., Bremner, A., Strother, S. et al. (2015). Can teachers accurately predict student performance? *Teaching and Teacher Education, 49*, 36–44. doi: 10.1016/j.tate.2015.01.012

Timmermans, A. C., de Boer, H., & van der Werf, M. P. C. (2016). An investigation of the relationship between teachers' expectations and teachers' perceptions of student attributes. *Social Psychology of Education, 19*, 217–240. doi: 10.1007/s11218-015-9326-6.

Trittel, M., Gerich, M., & Schmitz, B. (2014). Training prospective teachers in educational diagnostics. In S. Krolak-Schwerdt, S. Glock, & M. Böhmer (Eds.), *Teachers' professional development: Assessment, training, and learning* (pp. 63–78). Rotterdam: Sense.

Tze, V. M., Daniels, L. M., & Klassen, R. M. (2016). Evaluating the relationship between boredom and academic outcomes: A meta-analysis. *Educational Psychology Review, 28*, 119–144. doi: 10.1007/s10648-015-9301-y

Urhahne, D. (2015). Teacher behavior as a mediator of the relationship between teacher judgment and students' motivation and emotion. *Teaching and Teacher Education, 45*, 73–82. doi: 10.1016/j.tate.2014.09.006

Urhahne, D., Chao, S.-H., Florineth, M. L., Luttenberger, S., & Paechter, M. (2011). Academic self-concept, learning motivation, and test anxiety of the underestimated student. *British Journal of Educational Psychology, 81*, 161–177. doi: 10.1348/000709910X504500

Urhahne, D., Timm, O., Zhu, M., & Tang, M. (2013). Sind unterschätzte Schüler weniger leistungsmotiviert als überschätzte Schüler? [Are underestimated students less achievement motivated than overestimated students?] *Zeitschrift für Entwicklungspsychologie und Pädagogische Psychologie, 45*, 34–43. doi: 10.1026/0049-8637/a000079

Urhahne, D., Zhou, J., Stobbe, M., Chao, S.-H., Zhu, M., & Shi, J. (2010). Motivationale und affektive Merkmale unterschätzter Schüler. Ein Beitrag zur diagnostischen Kompetenz von Lehrkräften [Motivational and affective characteristics of underestimated students: A contribution to the diagnostic competence of teachers]. *Zeitschrift für Pädagogische Psychologie, 24*, 275–288. doi: 10.1024/1010-0652/a000021

Urhahne, D., & Zhu, M. (2015). Teacher judgment and student motivation. In C. M. Rubie-Davies, J. M. Stephens, & P. Watson (Eds.), *International handbook of social psychology of the classroom* (pp. 304–315). London, UK: Routledge.

Varner, G. F. (1923). Improvement in rating the intelligence of pupils. *The Journal of Educational Research, 8*, 220–232. doi: 10.1080/00220671.1923.10879399

von der Embse, N. P., Jester, D., Roy, D., & Post, J. (2018). Test anxiety effects, predictors, and correlates: a 30-year meta-analytic review. *Journal of Affective Disorders, 227*, 483–493. doi: 10.1016/j.jad.2017.11.048

Voyer, D., & Voyer, S. D. (2014). Gender differences in scholastic achievement: A meta-analysis. *Psychological Bulletin, 140*, 1174–1204. doi: 10.1037/a0036620

Weiner, B. (1980). *Human motivation*. New York: Holt, Rinehart & Winston.

Westphal, A., Kretschmann, J., Gronostaj, A., & Vock, M. (2018). More enjoyment, less anxiety and boredom: How achievement emotions relate to academic self-concept and teachers' diagnostic skills. *Learning and Individual Differences, 62*, 108–117. doi: 10.1016/j.lindif.2018.01.016

White, K. R. (1982). The relation between socioeconomic status and academic achievement. *Psychological Bulletin, 91*, 461–481. doi: 10.1037/0033-2909.91.3.461

Wijnia, L., Loyens, S. M.M., Derous, E., & Schmidt, H. G. (2016). University teacher judgments in problem-based learning: Their accuracy and reasoning. *Teaching and Teacher Education, 59*, 203–212. doi: 10.1016/j.tate.2016.06.005

Wright, D., & Wiese, M. J. (1988). Teacher judgment in student evaluation: A comparison of grading methods. *Journal of Educational Research, 82*, 10–14. doi: 10.1080/00220671.1988.10885858

Zayac, R. M., Ratkos, T., Frieder, J. E., & Paulk, A. (2016). A comparison of active student responding modalities in a general psychology course. *Teaching of Psychology, 43*, 43–47. doi: 10.1177/0098628315620879

Zeidner, M. (1998). *Test anxiety: The state of the art*. New York: Plenum.

Zhou, J., & Urhahne, D. (2013). Teacher judgment, student motivation, and the mediating effect of attributions. *European Journal of Psychology of Education, 28*, 275–295. doi: 10.1007/s10212-012-0114-9

Zhu, C., & Urhahne, D. (2018). The use of learner response systems in the classroom enhances teachers' judgment accuracy. *Learning and Instruction, 58*, 255–262. doi: 10.1016/j.learninstruc.2018.07.011

Zhu, M. (2014). *Accuracy of foreign language teacher judgment: Mechanisms and consequences* (Doctoral dissertation, Ludwig-Maximilians-Universität München).

Zhu, M., & Urhahne, D. (2014). Assessing teachers' judgements of students' academic motivation and emotions across two rating methods. *Educational Research and Evaluation, 20*, 411–427. doi: 10.1080/13803611.2014.964261

Zhu, M., & Urhahne, D. (2015). Teachers' judgements of students' foreign-language achievement. *European Journal of Psychology of Education, 30*, 21–39. doi: 10.1007/s10212-014-0225-6

Zhu, M., Urhahne, D., & Rubie-Davies, C. M. (2018). The longitudinal effects of teacher judgement and different teacher treatment on students' academic outcomes. *Educational Psychology, 38*, 648–668. doi: 10.1080/01443410.2017.1412399

Ziegler, A., Dresel, M., Schober, B., & Stöger, H. (2005). Ulm Motivational Test Battery (UMTB): Documentation of Items and Scales. (Ulm Educational Research Report, No. 15) Ulm: Ulm University, Department of Educational Psychology.

Ziegler, A., Dresel, M., & Stöger, H. (2008). Addresses of performance goals. *Journal of Educational Psychology, 100*, 643–654. doi: 10.1037/0022-0663.100.3.643

APPENDICES

Appendix A: Semi-Structured Questionnaire for Teachers' Judgment Strategies

Background information

Gender: ❒ male ❒ female

Subjects: ❒ English ❒ Mathematics ❒ Chinese

Class teacher: ❒ yes ❒ no

Age: ____ years

Teaching experience: ____ years

How many tests do you conduct in your class in one semester? ____ tests

How many days per week do the students get homework? ____ days

What percentage of class time is typically spent on each of the following activities?
Write a percentage for each activity. Write 0 (zero) if none. Please ensure that responses add up to 100%.

%	Administrative tasks (e.g., recording attendance, handing out school information or forms)
%	Keeping order in the classroom (maintaining discipline)
%	Actual teaching and learning
%	Correcting homework in the classroom
%	Other activities (e.g., talking with students after class or school)

In which way do you judge student achievement? Please mark the factors that you use and rank them. First mark them, then rank them.

		Rank the three factors
❒	Compare students among each other	
❒	Compare students with fixed criteria	
❒	Compare students to their prior achievement	

What **abilities and attitudes** of the students help you to judge their achievement? Please mark the factors that you use and rank them. First mark them, then rank the top three.

		Rank top three factors
❒	General Intelligence	
❒	Mathematical skills	
❒	Verbal skills	
❒	Learning strategies	
❒	Motivation	
❒	Interest and enjoyment	
❒	Self-confidence	
❒	Anxiety about the subject	
❒	other: _____	

What information about student behavior **during class** helps you to judge student achievement? Please mark the factors that you use and rank them. First mark them, then rank the top three.

		Rank top three factors
❒	raises hands often	
❒	communicates well with me	
❒	is prepared for class	
❒	concentrates well	
❒	has passion for the class	
❒	likes to join the classroom activities	
❒	other: _____	

What information from **homework** helps you to judge student achievement? Please mark the factors that you use and rank them. First mark them, then rank the top three.

		Rank the three factors
❒	finish homework on time	
❒	finish homework correctly	
❒	finish homework independently	
❒	other: _____	

What information from **tests** helps you to judge student achievement? Please mark the factors that you use and rank them. First mark them, then rank the top three.

		Rank top three factors
❒	Test performance	
❒	Test strategies	
❒	Grades of other subjects	
❒	Past academic records	
❒	Test anxiety	
❒	Student attribution of failure	
❒	other: _____	

What information from student behavior **after class** helps you to judge student achievement? Please mark the factors that you use and rank them. First mark them, then rank the top three.

		Rank top three factors
❒	Likes to ask you questions	
❒	Likes to help you	
❒	Likes to talk with you	
❒	Attends tutoring	
❒	Attends competitions	
❒	other: _____	

What information from **social interactions** helps you to judge student achievement? Please mark the factors that you use and rank them. First mark them, then rank the top three.

		Rank the three factors

❒	conversations with other teachers
❒	conversations with parents
❒	use of social media
❒	other: _____

What **demographic information** about the students helps you to judge student achievement? Please mark the factors that you use and rank them. First mark them, then rank the top three.

		Rank the three factors
❒	gender	
❒	age	
❒	physical appearance	
❒	parents' educational level	
❒	parents' economic status	
❒	other: _____	

Rank the different information fields to judge student achievement best.

	Rank the seven factors
Student abilities and attitudes	
Student behavior during class	
Student homework	
Student tests	
Student behavior after class	
Social interaction with other teachers or parents	
Demographic information	

How do you ensure that judgments about student achievement are soundly based?

_____

_____

_____

What do you do in case of uncertainties about judgments of student achievement?

_____

_____

_____

How many days of professional development do you attend within a year? ____ days

Have you ever attended a training related to judging student achievement? ❒ yes ❒ no

Would you like to take part in a training program on judging student achievement? ❒ yes ❒ no

Appendix B: The Classroom Response System Perceptions (CRiSP) Questionnaire

	Stimmt gar nicht	Stimmt eher nicht	Stimmt teils-teils	Stimmt eher	Stimmt genau
1. Durch die Benutzung von Clickern im Unterricht wurde zu viel Zeit vergeudet.	❒	❒	❒	❒	❒
2. Ich fände es gut, wenn im Unterricht weiterhin mit Clickern gearbeitet würde.	❒	❒	❒	❒	❒
3. Die Benutzung von Clickern hat den Wert des Unterrichts gesteigert.	❒	❒	❒	❒	❒
4. Durch die Clicker hat sich meine Motivation zum Lernen erhöht.	❒	❒	❒	❒	❒
5. Die Zusammenarbeit zwischen Schülern und Lehrer hat mit den Clickern gut funktioniert.	❒	❒	❒	❒	❒
6. Durch die Clicker habe ich sofortige Rückmeldungen bekommen, was ich weiß und was nicht.	❒	❒	❒	❒	❒
7. Die Benutzung der Clicker hat die Wahrnehmung der Meinungen und Einstellungen meiner Mitschüler verbessert.	❒	❒	❒	❒	❒
8. Clicker ermöglichen mir ein besseres Verständnis von Schlüsselbegriffen.	❒	❒	❒	❒	❒
9. Mein Lehrer hat die Ergebnisse der Clickerfragen genutzt, um das Verständnis der Klasse abzuschätzen und Stoff zu wiederholen, der noch verstanden wurde.	❒	❒	❒	❒	❒
10. Die Verwendung von Clickerfragen hat das Lernen des Faches verbessert.	❒	❒	❒	❒	❒
11. Ich glaube, dass Clicker mir mehr Kontrolle über mein Lernen geben als Lerneinheiten ohne Clicker.	❒	❒	❒	❒	❒
12. Der Gebrauch von Clickern hat mir geholfen, tiefer über den Lernstoff nachzudenken.	❒	❒	❒	❒	❒
13. Ich habe häufig die richtige Antwort gewählt, ohne es wirklich verstanden zu haben.	❒	❒	❒	❒	❒
14. Die Verwendung von Clickern hat mich selbstbewusst für die Teilnahme am Unterricht gemacht.	❒	❒	❒	❒	❒
15. Ich habe die Clicker meistens benutzt, wenn es angeboten wurde.	❒	❒	❒	❒	❒
16. Die Clicker haben die Häufigkeit meiner direkten Teilnahme am Unterricht erhöht.	❒	❒	❒	❒	❒
17. Die Verwendung von Clickern hat mir geholfen, aktiv am Unterricht teilzunehmen.	❒	❒	❒	❒	❒
18. Der Gebrauch von Clickern hat mir geholfen, im Unterricht aufmerksamer zu sein.	❒	❒	❒	❒	❒
19. Die Verwendung von Clickern hat meine Konzentration im Unterricht verbessert.	❒	❒	❒	❒	❒
20. Die Clicker haben dafür gesorgt, dass ich lieber am Unterricht teilgenommen habe.	❒	❒	❒	❒	❒
21. Für mich war der Gebrauch von Clickern als ein System zum Abstimmen leicht.	❒	❒	❒	❒	❒
22. Für mich war der Gebrauch von Clickern zu schwierig.	❒	❒	❒	❒	❒
23. Es war zu schwer zu verstehen was beim Gebrauch von Clickern von mir erwartet wurde.	❒	❒	❒	❒	❒
24. Es gab zu viele technische Probleme bei der Benutzung der Clicker.	❒	❒	❒	❒	❒

25. Die Verwendung von Clickern hat die Freude am Unterricht gesteigert.	❒	❒	❒	❒	❒
26. Andere Schüler konnten meine Antworten nicht sehen, was mich ermutigt hat, aktiv am Unterrichtteilzunehmen.	❒	❒	❒	❒	❒

Appendix C: Content of the Diary Items

1. My focus today was on correctly assessing the learning behavior of my students.

2. Today, I was motivated to get to the bottom of the causes of my students' learning difficulties.

3. Today, I had the feeling that I was able to assess the learning behavior of my students well.

4. Today, I took a thorough approach to assess the learning behavior of my students.

5. After today's class, I have reflected about whether I have assessed the behavior of my students properly.

6. In order to adequately assess the learning behavior of my students, I compared today their current learning behavior with their previous learning behavior.

7. I have reviewed my assessment of a student's current learning behavior in order to correct it if necessary.

8. Today, I have given a student constructive feedback on his or her learning behavior.

Appendix D: Standardized Mathematics Test (For Sixth Graders)

The test items were derived from the German Mathematical Test for Sixth Grade (Götz, L., Lingel, K., & Schneider, W. (2013). *DEMAT 6+*. Deutscher Mathematiktest für sechste Klassen. Göttingen: Hogrefe.)

Appendix E: Student Motivation and Emotion Questionnaires

Schülernummer (z. B. 6A_3): _____

Alter: _____ Klasse: _____

Mädchen ☐ Junge ☐

Wie soll angekreuzt werden?

Auf den folgenden Seiten stehen eine Reihe von Sätzen. Bitte lies die Aussagen im Fragebogen. Kreuze zu jeder Aussage immer nur ein Kästchen an und zwar dasjenige, welches deiner Meinung nach am besten zutrifft.

	Stimmt gar nicht	Stimmt eher nicht	Stimmt teils-teils	Stimmt eher	Stimmt genau
1. Ich gehöre in Mathe zu den guten Schülern.	❒	❒	❒	❒	❒
2. Mir fällt Mathe leicht.	❒	❒	❒	❒	❒
3. Die Hausaufgaben in Mathe sind für mich einfach.	❒	❒	❒	❒	❒
4. Wenn ich in Mathe dran komme, weiß ich die richtige Antwort.	❒	❒	❒	❒	❒
5. Es fällt mir leicht, in Mathe etwas zu verstehen.	❒	❒	❒	❒	❒
6. Ich bin gut in Mathe.	❒	❒	❒	❒	❒
7. Ich will jeden Tag in Mathematik etwas Neues lernen.	❒	❒	❒	❒	❒
8. Ich gebe in Mathematik mein Bestes.	❒	❒	❒	❒	❒
9. Ich strenge mich beim Lernen in Mathematik an.	❒	❒	❒	❒	❒
10. Ich versuche in Mathematik alles richtig zu machen.	❒	❒	❒	❒	❒
11. Ich versuche, auch ganz schwierige mathematische Aufgaben zu lösen.	❒	❒	❒	❒	❒
12. Ich strenge mich an, damit ich besser rechnen kann.	❒	❒	❒	❒	❒
13. Mathematiklernen macht mir Freude.	❒	❒	❒	❒	❒
14. Ich freue mich auf den Matheunterricht.	❒	❒	❒	❒	❒
15. Ich habe Lust, in Mathematik etwas zu lernen.	❒	❒	❒	❒	❒
16. Rechnen macht mir Spaß.	❒	❒	❒	❒	❒
17. Der Mathematikunterricht macht mir Spaß.	❒	❒	❒	❒	❒
18. Ich arbeite im Mathematikunterricht gern mit.	❒	❒	❒	❒	❒
19. Ich habe Angst, in Mathe schlechte Noten zu bekommen.	❒	❒	❒	❒	❒
20. In Mathe mache ich Fehler, weil ich Angst habe.	❒	❒	❒	❒	❒

21. Ich bin im Matheunterricht nervös.	❒	❒	❒	❒	❒
22. Ich habe Angst vor einer Matheprobe.	❒	❒	❒	❒	❒
23. Bei einer Matheprobe bin ich aufgeregt.	❒	❒	❒	❒	❒
24. Ich mache mir Sorgen, ob ich Mathe gut schaffen werde.	❒	❒	❒	❒	❒
25. Die Lösung schwieriger Mathematikprobleme gelingt mir immer, wenn ich mich darum bemühe.	❒	❒	❒	❒	❒
26. In Mathematik bin ich sicher, auch den schwierigsten Stoff zu verstehen.	❒	❒	❒	❒	❒
27. Ich bin überzeugt, dass ich auch die kompliziertesten Mathematikaufgaben lösen kann.	❒	❒	❒	❒	❒
28. Wenn in Mathematik ein Problem auftaucht, kann ich es aus eigener Kraft meistern.	❒	❒	❒	❒	❒
29. Ich bin überzeugt, dass ich alle Fertigkeiten, die zur Lösung von Mathematikproblemen gebraucht werden, erlernen und beherrschen kann.	❒	❒	❒	❒	❒
30. Für jedes mathematische Problem kann ich eine Lösung finden.	❒	❒	❒	❒	❒

Denk bitte an die nächste Probe in Mathematik!

Was denkst du, welche Note wirst du in der nächsten Schulaufgabe erhalten? _____

Mit welcher Note bei der nächsten Schulaufgabe wärst du gerade noch zufrieden? _____

Appendix F: Teacher Judgment Questionnaires

Schülernummer (Bitte nummerieren Sie in einer Klassenliste die Schülerinnen und Schüler fortlaufend durch, so dass Ihre Einschätzungen den Schülerfragebögen zugeordnet werden können)

Schülernummer _____

1. Wie viele der 25 Aufgaben des Mathematiktests löst der Schüler richtig?

_____ Aufgaben

2. Bitte schätzen Sie folgende Merkmale des Schülers im Vergleich zu anderen Schülern im selben Alter ein:

Fähigkeitsselbstkonzept (Wie schätzt der Schüler seine Fähigkeiten in Mathematik ein?)

Sehr viel geringer	Deutlich geringer	Geringer	Etwas geringer	Gleich	Etwas größer	Größer	Deutlich größer	Sehr viel größer
❍	❍	❍	❍	❍	❍	❍	❍	❍

Leistungsangst (Wie viel Angst hat der Schüler vor Mathematik?)

Sehr viel weniger	Deutlich weniger	Weniger	Etwas weniger	Gleich	Etwas mehr	Mehr	Deutlich mehr	Sehr Viel mehr
❍	❍	❍	❍	❍	❍	❍	❍	❍

Lernmotivation (Wie stark engagiert sich der Schüler beim Mathematiklernen?)

Sehr viel schwächer	Deutlich schwächer	Schwächer	Etwas Schwächer	Gleich	Etwas stärker	Stärker	Deutlich stärker	Sehr viel stärker
❍	❍	❍	❍	❍	❍	❍	❍	❍

Interesse (Wie sehr mag der Schüler das Fach Mathematik?)

Sehr viel weniger	Deutlich weniger	Weniger	Etwas weniger	Gleich	Etwas mehr	mehr	Deutlich mehr	Sehr viel mehr
❍	❍	❍	❍	❍	❍	❍	❍	❍

Selbstwirksamkeit (Wie sehr ist der Schüler überzeugt, ein guter Mathematik-Lernender zu sein?)

Sehr viel weniger	Deutlich weniger	Weniger	Etwas weniger	Gleich	Etwas mehr	mehr	Deutlich mehr	Sehr viel mehr
❍	❍	❍	❍	❍	❍	❍	❍	❍

Welche Note erwartet der Schüler für seine nächste Schulaufgabe in Mathematik? ____

Mit welcher Note wäre der Schüler in der nächsten Schulaufgabe in Mathematik gerade noch zufrieden?

____

Versicherung (gem. § 4 Abs. 3 Satz 1 Nr. 5 PromO):

Ich versichere hiermit

- an Eides statt, dass ich die Dissertation selbständig angefertigt, außer den im Schriftenverzeichnis sowie den Anmerkungen genannten Hilfsmitteln keine weiteren benutzt und die Herkunft der Stellen, die wörtlich oder sinngemäß aus anderen Werken übernommen sind, bezeichnet habe,
- dass ich die Dissertation nicht bereits in derselben oder einer ähnlichen Fassung an einer anderen Fakultät oder einer anderen Hochschule zur Erlangung eines akademischen Grades eingereicht habe.

…….

(Unterschrift)