THE FUTURE OF EDUCATION RESEARCH

Teachers' Professional Development

Assessment, Training, and Learning

Sabine Krolak-Schwerdt, Sabine Glock and Matthias Böhmer (Eds.)

SensePublishers

Teachers’ Professional Development

The Future of Education Research

Volume 03

Series editor

Daniel Tröhler
University of Luxembourg

One characteristic of modern societies is that they are likely to assign their social problems to education. Arising in the specific context of the late eighteenth century, this 'educational reflex' paved the way for education to become an important social factor on regional, national and global scales. Witnesses for this upswing are for instance the expansion of compulsory schooling, the state organization and tertiarization of teacher education and thus the introduction of education departments in the universities.

However, in contrast to the social artefact of modern societies – pluralism in languages, cultures, values, and customs –, education research seems in many respects still committed to ideas of unity or uniformity: For instance, the global standardization movement fosters uniformity in curriculum and content to serve the purpose of dominant global evaluation schemes, which in turn are based on the idea of human cognition as an immutable arrangement of mental processes with regard to learning. Moreover, critics of these developments often argue with arguments and convictions that can be traced back to the time when the education sciences emerged in the context of the cultural and political idea of the uniform national state.

Obviously, today's education research often operates using concepts that are derived from ideas of unity and uniformity in order to tackle the challenges of cultural and linguistic plurality in the context of democratic societies. This is both a paradox and an occasion to reflect upon the present and future role of education research in the context of modern societies in three attempts: *Education Systems in Historical, Cultural, and Sociological Perspectives* (Vol. 1); *Multilingualism and Multimodality: Current Challenges for Educational Studies* (Vol. 2); *Teachers' Professional Development: Assessment, Training, and Learning* (Vol. 3).

Teachers' Professional Development

Assessment, Training, and Learning

Edited by

Sabine Krolak-Schwerdt
University of Luxembourg, Luxembourg
Sabine Glock
University of Luxembourg, Luxembourg

and

Matthias Böhmer
University of Luxembourg, Luxembourg

SENSE PUBLISHERS
ROTTERDAM/BOSTON/TAIPEI

A C.I.P. record for this book is available from the Library of Congress.

ISBN: 978-94-6209-534-2 (paperback)
ISBN: 978-94-6209-535-9 (hardback)
ISBN: 978-94-6209-536-6 (e-book)

Published by: Sense Publishers,
P.O. Box 21858,
3001 AW Rotterdam,
The Netherlands
https://www.sensepublishers.com/

Printed on acid-free paper

TABLE OF CONTENTS

DANIEL TRÖHLER

THE FUTURE OF EDUCATION RESEARCH

Introduction to the Series of Three Volumes

One characteristic of modern societies is that they are likely to assign their social problems to education. Arising in the specific context of the late eighteenth century, this 'educational reflex' paved the way for education to become an important social factor on local, regional, national and global scales. Witnesses for this upswing are, for instance, the expansion of compulsory schooling, the state organization and tertiarization of teacher education and thus the introduction of educational departments in the universities, the introduction of certificates for both students and teachers.

However, in contrast to the social artefact of modern societies – pluralism in languages, cultures, values, and customs – the educational sciences seem in many respects still committed to ideas of unity or uniformity: For instance, the global standardization movement fosters uniformity in curriculum and content to serve dominant global evaluation schemes. These schemes in turn are based on the idea of human cognition as an immutable arrangement of mental processes with regard to learning. And the critics of these developments often argue with motives, arguments, and convictions that can be traced back to the time when the educational sciences emerged in the context of the cultural and political idea of the uniform (and of course superior) national state. In other words: Today, often the education sciences operate using concepts that are derived from ideas of unity and uniformity in order to tackle the challenges of cultural and linguistic plurality in the context of democratic societies. This obviously is both a paradox and an occasion to reflect about the present and future role of the educational sciences in the context of modern societies.

With over 40% of inhabitants not having Luxembourgish passports, Luxembourg is a multinational and thus a multilingual and multicultural society. With its three official languages Luxembourgish, German, French, and with Portuguese as first language of nearly 20% of the inhabitants, it is also a multilingual society. Against this background, Luxembourg is predestined to evaluate the 'educational reflex' mentioned above, the assigning of social problems to education. The University of Luxembourg responded to this desideratum by making 'Education and Learning in Multilingual and Multicultural Contexts' a Research Priority in the frame of the current four-year-plan (2010-2013).

One particular challenge of this research priority is the self-reflection or critical self-evaluation of the educational sciences in the context of the social expectations concerning education. Therefore, one of the major aims of "Education and Learning in Multilingual and Multicultural Contexts" was to assess the future of educational research with outstanding international scholars. The 2010-2013 lecture series "The Future of Education Research" is an integral part of this research priority. Here the international discussion is not restricted to questions regarding technical feasibility and methods of educational ambitions. Self-reflection or critical self-evaluation meant precisely refraining from compliant adoptions of research desiderata defined by stakeholders of political, cultural, religious, or developmental institutions and being engaged in the (self-) critical assessment of the legitimacy and general feasibility of educational desiderata, that is of social expectations emerging from the educational reflex. Education research was defined not simply as a service towards fulfilling social expectations but like any other academic discipline a field in which its actors, the researchers, define the appropriateness of its research agenda – research questions and methods – in the realm of their peers.

With these premises, the future of education research is defined to be international, self-reflexive, and interdisciplinary and to include a broad range of traditional academic disciplines such as the education sciences in the narrower sense, psychology, sociology, linguistics, history, political sciences, cognitive sciences, and neurology sciences. And it is meant to focus on the macro, meso and micro levels of education questions and problems analytically, empirically, and historically. The invited international colleagues addressed their respective scholarship to the topic under consideration, the future of education research, in one of three lecture series at the University of Luxembourg from 2010 to 2013. In accordance with the interdisciplinary approach, the relevant questions were not clustered around traditional disciplines but around several focal points, resulting in this series of the following three volumes to be published between 2011 and 2014:

- *Education Systems in Historical, Cultural, and Sociological Perspectives* (Vol. 1)
- *Multimodality and Multilingualism: Current Challenges for Educational Studies* (Vol. 2)
- *Teachers' Professional Development: Assessment, Training, and Learning* (Vol. 3)

We greatly appreciate the support of the University of Luxembourg and extend thanks for the opportunity to establish a Research Priority dedicated to "Education and Learning in Multilingual and Multicultural Contexts," within which the lecture series "The Future of Education Research" is being held. We are grateful to all the excellent international scholars participating in this research discussion. And last but not least, we sincerely thank Peter de Liefde of Sense Publishers for his support of this series and for giving us, by means of publication, the opportunity to open up this discussion on a more global level.

Walferdange, Luxembourg, August 2011

Daniel Tröhler, Head of the Research Priority
"Education and Learning in Multilingual and Multicultural Contexts",
University of Luxembourg

SABINE KROLAK-SCHWERDT, SABINE GLOCK &
MATTHIAS BÖHMER

INTRODUCTION

The conditions and consequences of societal change are the focal point of current debates concerning professional development of actors in the educational domain, most notably teachers. The major goal is to make teacher education a profession with a research base and formal body of knowledge and to ensure that teachers are fully prepared in accordance with professional standards. In recent years, the teaching profession has begun to identify and develop the knowledge base that will frame the education curriculum.

A central aspect of teachers' professional knowledge and competence is the ability to assess students' achievements adequately. Giving grades and marks is one prototypical task in this context. Besides giving grades, assessments for school placements or tracking decisions belong to these tasks. Relevant students' characteristics which influence teachers' assessments do not only involve academic achievement but also students' responses to different task demands as well as non-academic characteristics such as learning motivation or school anxiety. Teachers' assessments have substantial relevance for individual students, and consequently, high competence in assessing students correctly is seen as a key skill for teachers.

Closely associated with the investigation of teachers' assessment competences and, more specifically, the investigation of conditions associated with high quality of assessments is the development and evaluation of teacher training programs to improve professional competences. In recent years, there has been considerable progress in the domain of professional teacher training; however, only a very limited number of studies are dedicated to the question to what extend training programs might offer valuable approaches to improve the quality of assessments and to implement high assessment competences.

Another important field which is closely related to teachers' competences concerns the question how teachers' professional development is linked to students' learning and learning outcomes. In recent years, the societal demand for evidence that teachers' professional development will result in improved student learning outcomes is increasing. Current theorizing postulates a long chain of intermediate steps and variables which links teachers' professional development to students' learning. For instance, teachers' beliefs about (good) teaching methods and students' beliefs about learning might constitute such intermediate variables. There is, however, little research which covers the whole causal chain.

S. Krolak-Schwerdt et al., (Eds.), Teachers' Professional Development, 1–4.

Taken together, questions on assessment, training, and learning in the professional development of teachers have not been fully discussed. The identification of these research gaps was the reason for dedicating the third round of lectures in the University of Luxembourg's 2010-2013 lecture series "The Future of Educational Research" to the topic of professionalization of teachers in these domains. It was therefore our privilege to invite outstanding international scholars in different academic disciplines to present ideas about open research questions concerning the domains of assessment, training, and learning in the professional development of teachers.

In correspondence to these thematic foci, the first part of this volume is concerned with teachers' assessment competences. Conceptualizing assessments as judgments about students' characteristics, recent approaches frequently take a view on the quality of assessments as "judgment accuracy". Correspondingly, the first part of this volume is concerned with teachers' judgments of students' academic achievement and judgment accuracy.

Südkamp, Kaiser, and Möller present a comprehensive model of teacher-based judgments of students' academic achievement. In line with the model, the authors present the results of a meta-analysis of 75 field studies reporting correlational data on the relationship between teachers' judgments of students' academic achievement and students' performance on a standardized achievement test. As to teacher characteristics, main results are that teachers who are informed about the content of a test prior to their judgment of students' academic achievement perform better than uninformed teachers. Moreover they found that the congruence between the teachers' judgment task and the achievement test administered to students is related to teacher judgment accuracy, with higher congruence being associated with higher accuracy levels. To further analyze the causal role of these variables, the authors recommend the combination of two research approaches, that is, an experimental approach and the validation of its results by field studies.

Artelt and Rausch review the empirical findings on teachers' judgment accuracy and discuss potential moderators. Beyond students' characteristics, task characteristics contribute to teachers' accuracy, as teachers' judgment accuracy is increasing with the correspondence between (1) the judgment scale and the test scale and (2) the judgment domain and the test domain. Artelt and Rausch present findings from the BiKS research group in Bamberg and provide evidence for teachers' global achievement judgments of a particular domain being more accurate than their task-specific judgments. Task specific judgments relate to students' ability to solve particular items in a test, while global judgments have no relations to particular tests. In the remainder of the chapter, the authors discuss necessary conditions for high judgment accuracy to occur. Among other variables, teachers' stereotypical expectations contribute to judgment accuracy.

Pit-ten Cate, Krolak-Schwerdt, Glock, and Markova focus on the change of cognitive processes in order to overcome expectation biases and to improve teachers' judgment accuracy. Frequently, the use of stereotypical knowledge about students is discussed as a potential source of judgment biases. However, the use of stereotypical knowledge in judgments is not inevitable, as there are techniques to overcome stereotypical biases such as the training of stereotype suppression. Furthermore, the motivation of the person making a judgment also plays a pivotal role in preventing teachers from reliance on stereotypical knowledge. It is demonstrated that one possible mechanism to increase accuracy motivations among teachers is to make them highly accountable for their judgment. A third aspect to increase teachers' judgment accuracy consists of using of statistical prediction rules for judgment formation. These are formal decision rules on how information which has proven to be diagnostic for the judgment task should be weighted and integrated into a judgment.

After having discussed the conditions necessary for sufficient teachers' judgment accuracy and how judgment accuracy could be improved, the second part of this volume deals with the development and evaluation of teacher training programs. Trittel, Gerich, and Schmitz are concerned with the evaluation of a program developed to train prospective teachers in assessment competence. They introduce a hands-on seminar which is based on a process model of teachers' assessment competence. This program provides prospective teachers with theoretical knowledge about educational assessments and corresponding quality criteria as well as with knowledge about judgment biases. Additionally, prospective teachers are introduced to assessment instruments and methods. The application of these instruments and methods is practiced in lessons where prospective teachers are also prepared to plan supportive measures relying on the preceding assessment process. First results demonstrate the usefulness of this program in training prospective teachers' assessment competence.

In the paper of Vermunt, teacher education and professional development is suggested to be causally related to student learning outcomes. Drawing on this chain model of teacher education and student learning outcomes Vermunt discusses the different components of student learning as well as innovative teaching-learning methods. The application of new teaching-learning methods such as problem-based teaching has proven to foster active student learning and to increase students' self-regulation. These new methods do not only challenge students' learning capacity but also teachers' professional development and learning. Unlike traditional teaching methods which require teachers to explain the subject matter, innovative teaching-learning methods require teachers to fulfill different roles depending on the method applied. Thus, teacher learning and the investigation of factors contributing to teachers' professional development are of high importance. Vermunt reviews factors which have been empirically shown to be related to teachers' learning and discusses issues such as the development of adequate measures of teacher learning to be addressed by future research.

Richter, Kunter, Klusmann, Lüdke, and Baumert examine teachers' uptake of formal and informal learning opportunities across the teaching career by use of data from 1939 German secondary teachers in 198 schools. Results show that formal learning opportunities, that is, in-service training, are used most frequently by mid-career teachers, whereas informal learning opportunities exhibit distinct patterns across the teaching career. Specifically, the use of professional literature increases with teacher age, but teacher collaboration decreases. Teachers' work engagement and professional responsibilities are hypothesized to predict changes over the career. These variables partly predict uptake of learning opportunities, but in contrast to the hypothesis, they do not fully explain the age-related differences observed.

The final two contributions to this volume are more specific in their respective topic. Bromme, Pieschl, and Stahl are concerned with student learning. More specifically, they investigate how epistemological beliefs of students, that is, students' beliefs about the nature of knowledge, affect their learning. It is assumed that more sophisticated beliefs are associated with better alignment between task complexity and the learning process. Participants in their study are biology and humanity students who were given a list of tasks of different complexity. For each task, they completed a questionnaire. The results show that there is a reasonable alignment in that students' answers to the questionnaire are related to task complexity in a meaningful way. Furthermore, alignment is linked to epistemological beliefs. For example, students with sophisticated beliefs judge the use of deep processing learning strategies more important across all tasks.

The paper of Klapproth and Schaltz is concerned with the predictive validity of academic and vocational-training achievement. The authors present a review of 52 studies in which the quality of predictors of academic achievements are investigated at the level of primary school, secondary school as well as achievements in universities and/or vocational training programs. A general finding is that achievements at each level can be predicted with sufficient accuracy by the achievements at the preceding school level. However, there are large differences in that shorter time periods of prediction, standardized achievement tests (as compared to unstandardized measures) and achievements in secondary school (as compared to other levels) have a higher accuracy of prediction.

We are grateful to all colleagues for their contribution to this volume. We hope that this volume will offer an important perspective on the domains of assessment, training, and learning in the professional development of teachers.

AFFILIATIONS

Sabine Krolak-Schwerdt
University of Luxembourg

Sabine Glock
University of Luxembourg

Matthias Böhmer
University of Luxembourg

ANNA SÜDKAMP, JOHANNA KAISER & JENS MÖLLER

TEACHERS' JUDGMENTS OF STUDENTS' ACADEMIC ACHIEVEMENT

Results From Field and Experimental Studies

INTRODUCTION

Teacher judgments of student achievement have a considerable impact on students' learning experiences and educational trajectories. Also, many instructional decisions are determined by teachers' subjective judgments of their students' achievement. The ability to accurately gauge student outcomes is therefore one of the key characteristics of a good teacher.

In the first part of this chapter, we provide a comprehensive description of our heuristic model of teacher judgment accuracy, which was first introduced in our meta-analysis on this issue (Südkamp, Kaiser, & Möller, 2012). In addition, we summarize the key findings of the meta-analysis. Studies included in the meta-analysis were limited to field studies. In the second part of this chapter, we introduce an experimental approach to the study of teacher judgment accuracy. In our own empirical research, we used the Simulated Classroom, which is a computer simulation of a classroom situation. Here, factors relevant to real classroom situations (e.g., student achievement, motivation, gender, subject, number of students, lesson length, and content covered) can be experimentally manipulated. Again, we provide key empirical findings on teacher judgment accuracy as well as on potential moderators. Finally, advantages and disadvantages of the field and experimental approach are discussed.

A MODEL OF TEACHER JUDGMENT ACCURACY

Given the important implications of teacher judgments (see Artelt, 2013; in this book), the question of their accuracy is critical. Accurate assessment of students' performance is a necessary condition for teachers being able to adapt their instructional practices, to make fair placement decisions, and to support the development of an appropriate academic self-concept. In order to arrange possible influences on teacher judgment accuracy systematically, we propose a heuristic model of teacher judgment accuracy, which is displayed in Figure 1.

Teacher judgment accuracy is at the core of the model. It represents the correspondence between teachers' judgments of students' academic achievement and

S. Krolak-Schwerdt et al., (Eds.), Teachers' Professional Development, 5–25.

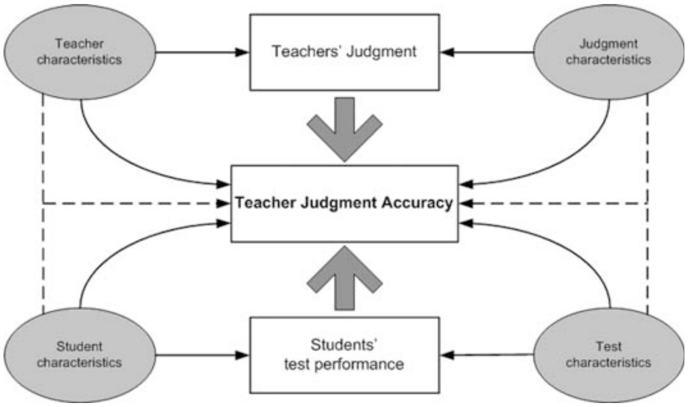

Figure 1. Heuristic model of teacher judgment accuracy.

students' actual academic achievement measured by a standardized test. Usually, the correlation between the two is used as a measure of this correspondence. However, other indicators, such as the average difference between teacher judgments and students' actual performance, can also be used.

A student's test performance is the result he or she achieves on an academic achievement test. On the one hand, this result may depend on student characteristics such as prior knowledge, motivation, and intelligence. On the other hand, it may depend on test characteristics such as subject area, the specific tasks set, or task difficulty. In turn, a teacher's judgment may depend on teacher characteristics such as professional expertise or stereotypes about students, and/or on judgment characteristics (e.g., whether the teacher is asked to judge a specific student competency, such as oral reading fluency, or to provide a global judgment of academic ability).

According to our model, teacher judgment accuracy is also influenced by the correspondence between judgment characteristics and test characteristics (dashed line). Potentially, the achievement test may measure a very specific academic ability (e.g., arithmetic skills), whereas the focus of the teachers' judgment task is broader (e.g., rating students' overall ability in mathematics), making it more difficult for teachers to make an accurate judgment. Another relationship that may influence teacher judgment accuracy is the correspondence between teacher characteristics and student characteristics (e.g., gender, ethnicity).

In the next section, we give a more detailed description of the model, reporting key empirical results for each aspect. We first summarize research findings on teacher judgment accuracy and then describe how teacher characteristics, judgment characteristics, student characteristics, and test characteristics influence teacher judgment accuracy.

Teacher Judgment Accuracy

Most research on teacher judgment accuracy examines the relationship between teachers' judgments of students' achievement and students' actual performance on measures of achievement in various subject areas. Although, in most studies, academic achievement is measured by a standardized achievement test, some studies have used self-constructed, less standardized tests (e.g., curriculum-based measurement [CBM] procedures, see below).

The most commonly reported measure quantifying the correspondence between teachers' judgments and students' actual achievement is the correlation between the two. Overall, moderate to high correlations are reported (Begeny et al., 2008; Demaray & Elliot, 1998; Feinberg & Shapiro, 2003). For example, Feinberg and Shapiro (2009) reported correlations of .59 and .60 between teachers' judgments and students' decoding skills and reading comprehension, as measured by subtests of the Woodcock-Johnson-III-Test of Achievement. In the same study, a correlation of .64 was found between students' oral reading fluency as measured by a CBM procedure and teachers' predictions of oral reading fluency. In a review of 16 studies, Hoge and Coladarci (1989) found a median correlation of .66 between teachers' judgments and students' achievement on a standardized test. In our recent meta-analysis on teacher judgment accuracy (Südkamp et al., 2012), the overall mean effect size was found to be .63. On the one hand, these results may be interpreted as indicating that teachers' judgments are quite accurate; on the other hand, their judgments are evidently far from perfect, and more than two thirds of the variance in teachers' judgments cannot be explained by student performance. Additionally, the correlations found ranged substantially across studies, from .28 to .92 (Hoge & Coladarci, 1989) and from -.03 to .84 (Südkamp et al., 2012)

Methodological differences between studies need to be taken into account when considering the differences in results across studies. We therefore focus on how teacher judgment accuracy relates to judgment and test characteristics as well as to teacher and student characteristics, as discussed in detail below.

Several authors have noted the problems of relying solely on the correlation between teachers' judgments and students' performance on standardized achievement tests. For example, if teachers systematically perceive their students to be more or less competent than indicated by their performance on objective measures, their judgments may still be highly correlated with students' performance (Eckert, Dunn, Codding, Begeny, & Kleinmann, 2006; Feinberg & Shapiro, 2003; Graney, 2008).

Indeed, the findings of studies using indicators other than correlations as measures of teacher judgment accuracy suggest that teacher judgments are rather inaccurate. Bates and Nettelbeck (2001) subtracted students' reading accuracy and reading comprehension scores on a standardized achievement test from teachers' predictions of these scores. Teachers generally overestimated the performance of the six- to eight-year-old students; inspection of the difference scores revealed that this held to a greater extent for low-achieving readers than for average- and high-achieving students. In a study by Eckert et al. (2006), CBM material was used as an indicator of students' mathematics and reading skills. Teachers were asked to estimate students' reading and mathematics level (mastery, instructional, or frustrational). This judgment was compared with students' actual reading and mathematics level as measured by the CBM material via percentage agreement. The results indicated that teachers overestimated students' performance across most mathematics skills and on reading material that was at or below grade level. In line with this result, Begeny et al. (2008) found that teachers' judgments of students with average to low oral reading fluency scores were rather inaccurate, and Feinberg and Shapiro (2003) reported that teachers generally overestimated the performance of low-achieving readers.

Nevertheless, studies have revealed large interindividual differences in teachers' ability to judge student performance accurately (Helmke & Schrader, 1987). For example, Lorenz and Artelt (2009) reported moderate average correlations between teacher judgments and student performance in reading and mathematics for a sample of 127 teachers. The standard deviation for the mean of the correlations was .30 for reading and .39 for mathematics. Some teachers showed very high judgment accuracy; others, very low judgment accuracy. These findings raise the question of which individual teacher characteristics are related to teacher judgment accuracy.

Teachers' Judgments

As shown in our model, teachers' judgments are thought to depend on individual teacher characteristics and judgment characteristics. In the following, we first consider teacher characteristics that have been associated with teacher judgment accuracy and then discuss the influence of judgment characteristics on teacher judgment accuracy. A teacher's characteristics are thought to influence her or his judgment at various stages of the judgment process (e.g., reception, perception, interpretation). Teacher characteristics such as job experience (Impara & Plake, 1998), beliefs (Shavelson & Stern, 1981), professional goals (Schrader & Helmke, 2001), and teaching philosophy (Hoge & Coladarci, 1989) have been associated with teachers' judgment processes in the literature. Although the variability in the accuracy of teachers' judgments is well documented (Hoge & Coladarci, 1989; Helmke & Schrader, 1987; Südkamp et al., 2012), empirical research has not yet pinpointed teacher characteristics that influence judgment accuracy.

Teacher Characteristics

Job experience and exposure to students. Impara and Plake (1998) expected teachers with more years of teaching experience to be experts in judging the item difficulties of a science test. However, they found no relationship between years of science teaching experience and the difference in actual and predicted item performance. The studies by Demaray and Elliott (1998) and Mulholland and

Berliner (1992) substantiated these findings. As already noted by Hoge and Coladarci (1989), research on teacher judgment accuracy provides few clues as to whether differences in teacher judgment accuracy are attributable to teacher characteristics. In our meta-analysis on teacher judgment accuracy, we considered the main teacher characteristics that have been associated with teacher judgment accuracy. We expected years of teaching experience and length of exposure to the students rated (i.e., how long teachers had taught those students) to be positively related to teacher judgment accuracy.

Teachers' gender, and age. In addition, we probed for effects of teacher age and gender on teacher judgment accuracy. As teacher characteristics had only been examined in a small number of studies at the time we conducted our meta-analysis, we were not able to study their effects within our analyses.

Teachers' judgments also depend on the characteristics of the judgment made. Usually, these characteristics reflect methodological decisions been made by the authors of a study. For example, some studies (e.g., Demaray & Elliott, 1998) ask teachers to directly judge students' performance (e.g., to estimate the number of correct responses on an achievement test for each student), whereas others ask teachers to rate students' overall academic ability indirectly on a rating scale. These differences are summarized under the label "judgment characteristics" in our model. It can be assumed that different judgment characteristics affect the correspondence between teachers' judgments and students' academic achievement.

Judgment Characteristics

Direct versus indirect teacher judgments. According to Hoge and Coladarci (1989), a distinction must be made between direct and indirect teacher judgments. In some studies, teachers are asked to assess students' academic achievement on a standardized achievement test by estimating the number of items each student will solve correctly (Helmke & Schrader, 1987). This approach can be considered a direct rating. In other studies, teachers are asked to rate students' performance in a certain subject on a Likert-type rating scale (e.g., a 5-point rating scale; DuPaul, Rapport, & Perriello, 1991). Hoge and Coladarci (1989) refer to this type of approach as indirect rating. In line with the results of Hoge and Coladarci, Feinberg and Shapiro (2003, 2009) and Demaray and Elliott (1998) found higher correlations for direct teacher judgments than for indirect teacher judgments. For example, Feinberg and Shapiro (2003) found a correlation of r = .70 between students' test performance and direct teacher judgments, whereas the correlation with indirect teacher judgments was r = .62.

Points on the rating scale. Studies using rating scales to obtain teacher judgments differ in terms of the number of points on the rating scales implemented. Rating scales with many categories permit a sophisticated judgment, whereas scales with fewer categories allow a more global judgment. Generally, slightly higher correlations with students' actual performance are obtained for more sophisticated judgments than for more global judgments. To date, this variable has been neglected in empirical research on teacher judgment accuracy.

Judgment specificity. According to Hoge and Coladarci (1989), the distinction between direct and indirect teacher judgments also has implications for the specificity of the judgment. In general, direct judgments are more specific than are indirect judgments, as they are explicitly tied to the criterion in the judgment process. Indirect teacher judgments may be differentiated in terms of their degree of specificity. Following the approach used by Hoge and Coladarci, teachers' judgments can be allocated to one of five categories, ranging from low to high specificity. First, a judgment that requires teachers to rate students' academic achievement on a rating scale (e.g., poor – excellent) is considered to be of low specificity. Second, in a ranking, the teacher's task is to put the students of his or her class into rank order according to their achievement. Third, tasks requiring teachers to find grade equivalents for students' performance on a standardized achievement test are considered to be of average specificity. Fourth, tasks asking teachers to estimate the number of correct responses achieved by a student on a standardized achievement test are slightly less specific than the fifth and most specific category, in which teachers indicate students' item responses on each item of an achievement test. In their review, Hoge and Coladarci found a median correlation of .61 for studies using ratings, which was the predominant approach. The median correlations for studies using rank ordering (median r = .76), grade equivalents (median r = .70), number of correct responses (r = .67, for a single study), and item-based judgments (median r = .70) were indeed higher.

Norm-referenced vs. peer-independent judgments. In addition, teacher judgments may differ in whether they are norm-referenced or peer-independent. For example, Helwig et al. (2001) asked teachers to rate students' academic achievement on an absolute scale (very low proficiency – very high proficiency), whereas Hecht and Greenfield (2002) asked teachers to estimate students' academic achievement in relation to other members of the class (in the bottom 10% of the class – in the top 10% of the class). Hoge and Coladarci (1989) considered this aspect in their meta-analysis, but found no substantial difference between correlations. The median correlation for norm-referenced judgments was .68; that for peer-independent judgments was .64.

Domain specificity. Finally, teacher judgments differ in terms of their domain specificity. Whereas some studies ask teachers to judge students on a very specific ability (e.g., arithmetic skills; Karing, 2009), others ask them to judge students' overall academic achievement (e.g., Li, Pfeiffer, Petscher, Kumtepe, & Mo, 2008). To our knowledge, no studies to date have examined the influence of the domain specificity of teachers' judgments on teacher judgment accuracy. However, it seems

reasonable to hypothesize that it is easier to make a focused judgment on a domain-specific ability than to judge a student's overall academic ability.
Instead of distinguishing between direct and indirect teacher judgments, we used a slightly different categorization in our own meta-analysis. We distinguished between informed versus uninformed teacher judgments, considering whether teachers were informed or uninformed about the content of the test prior to rating students' academic achievement and found significantly higher correlations between teachers' judgments and students' test performance for informed than for uninformed teacher judgments. No effects of the other judgment characteristics (i.e., number of points on rating scales, judgment specificity, norm-referenced versus peer-independent judgments) were found.

Students' Test Performance

Teacher judgment accuracy also depends on students' test performance, which in turn depends on individual student characteristics and characteristics of the test.

Student Characteristics

Several student characteristics have been identified as influencing the accuracy of teachers' judgments. For example, Bennett, Gottesman, Rock, and Cerullo (1993) found that teachers who perceived their students to exhibit bad behavior also perceived these students as low academic performers, regardless of the students' academic skills. In a study by Hurwitz, Elliott, and Braden (2007), the accuracy of teachers' judgments was related to students' disability status: Teachers predicted the mathematics test performance of students without disabilities more accurately than that of students with disabilities. Ritts, Patterson, and Tubbs (1992) found that teachers also took student attractiveness into account when judging students' academic achievement.

Gender. Although many studies have examined student gender as a potential moderator of teacher judgments, no consistent significant effects of student gender on teachers' judgments have been established (Demaray & Elliott, 1998; Hoge & Butcher, 1984). A large body of research in this area focuses on teacher expectations rather than teacher judgment accuracy. However, there is considerable overlap between the methodologies used. In both types of studies, teachers are asked to judge students' academic performance. Whereas teacher expectancy research tends to focus on differences in teachers' judgments of different groups (e.g., European vs. African American students, girls vs. boys), research on teacher judgment accuracy puts less emphasis on distinguishing between groups of students on the basis of a certain characteristic, focusing rather on the correspondence of teacher judgments with students' actual academic achievement. In teacher expectancy research, differences in teacher judgments of different groups are sometimes, but not always, controlled for students' actual academic achievement. Nevertheless, teacher expectancy research holds important implications for our meta-analysis, as it reveals factors influencing teachers' judgment. In their meta-analysis of teacher expectancy effects, Dusek and Joseph (1983) reviewed 16 studies in which student gender was related to a measure of teacher expectations of students' academic performance. The results indicated that student gender does not affect teacher expectancies of general academic achievement. Jussim and Eccles (1995) reported that teachers perceived girls in fifth grade mathematics classes as performing slightly higher than boys, but this perception was accurate, as girls in their study slightly outperformed boys on standardized achievement tests. In contrast, Tiedemann (2000) reported that teachers rated the mathematical ability of third and fourth grade boys to be slightly higher than that of their female counterparts, although these girls' and boys' grades in the previous school year were not significantly different. Controlling for first grade students' performance on a standardized reading test, Hinnant, O'Brien, and Ghazarian (2009) found that teachers tended to overestimate the reading ability of girls and to underestimate that of boys. In sum, there is only little evidence that teacher judgment accuracy is influenced by student gender.

Grade level. Studies have also examined whether teachers' judgment accuracy differs across grades. Kenny and Chekaluk (1993) reported that teachers' assignment of second grade children to reading categories (advanced, average, poor) was more accurate than their assignment of first graders and kindergarteners. In contrast, Maguin and Loeber (1996) reported significantly higher correlations between teacher judgments and students' reading and mathematics achievement for first graders than for fourth and seventh graders. Kuklinski and Weinstein (2001) reported that teacher expectancies accentuated achievement differences to a greater extent in the early elementary grades than in the later elementary grades.

Ethnicity. Research has shown that teacher expectancies of students are influenced by students' ethnicity (Baron, Tom, & Cooper, 1985; Dusek & Joseph, 1983). A recent meta-analysis of studies conducted in the United States (Tenenbaum & Ruck, 2007) revealed that teachers had more positive expectations for European American children than for ethnic minority children. In turn, Chang and Sue (2003) found that teachers had more positive expectations for Asian American students than for other ethnic groups. Wigfield, Galper, Denton, and Seefeldt (1999) showed that teachers' beliefs about students' academic abilities significantly predicted students' performance on a standardized test, and that teachers' ratings differed significantly between ethnic groups, with European American children being rated significantly higher than Hispanic children or African American children.

Socioeconomic background. The literature on teacher expectancies has identified students' social background as a factor that crucially informs teacher expectancies (Alexander, Entwisle, & Thompson, 1987; Jussim, Eccles, & Madon, 1996).

Alvidrez and Weinstein (1999) reported that teachers judged children from higher socioeconomic backgrounds more positively and students from lower socioeconomic backgrounds more negatively than the students' performance on the Wechsler Intelligence Scales would predict. Kennedy (1995) found the proportion of low-income students at a school to be strongly negatively correlated with teachers' perceptions of students' ability. In contrast to their expectations, Wigfield et al. (1999) did not find teachers' judgments to differ between former Head Start children from socioeconomically disadvantaged families and non-Head Start children. Likewise, Jussim and Eccles (1995) found no evidence that teachers judged students with lower socioeconomic status any less favorably than students with higher socioeconomic status. Hinnant et al. (2009) argued that teachers' expectations may influence the reading performance of minority groups in particular. In their study, first grade teachers' expectations were reliably linked to the third grade performance of minority boys, but not of White students or Non-White girls. In mathematics, teacher expectations were significantly and positively related to the later mathematics performance of children from families with low or average income, but unrelated to that of high-income children. In a study with a sample of kindergarten children from various ethnic groups living in low-income families, Hauser-Cram, Sirin, and Stipek (2003) found that teachers rated the children as less competent if they perceived value differences between themselves and the parents.

As is the case for teacher characteristics, however, few studies to date have reported information on the student sample. Moreover, any data available are not readily comparable across studies (e.g., only the percentage of female/male students was reported). Therefore, we decided not to conduct moderator analyses on student characteristics in this meta-analysis.

Test Characteristics

As shown in our model, students' test performance also depends on characteristics of the test. Like the judgment characteristics summarized above, these test characteristics in turn depend on methodological decisions made by the author(s) of the studies. In studies on teacher judgment accuracy, various instruments are used to measure students' academic achievement, ranging from highly specific tests measuring, for example, receptive vocabulary (e.g., the Peabody Picture Vocabulary Test used by Fletcher, Tannock, & Bishop, 2001) to broader tests measuring students' performance in different subject areas (e.g., the Kaufman Test of Academic Achievement measuring achievement in mathematics, reading, and spelling used by Demaray & Elliott, 1998). Our model summarizes such differences between tests under the label "test characteristics." Various test characteristics can be assumed to influence the correspondence between teachers' judgments and students' performance.

Subject matter. Comparing correlations between teachers' judgments and students' academic achievement in different subjects, Hopkins, George, and Williams (1985) found that correlations were significantly lower for social studies and science than for language arts, reading, and mathematics. Using CBM procedures to gauge students' academic achievement, Eckert et al. (2006) found higher correlations for reading than for mathematics. In turn, Coladarci (1986) reported teachers' judgments to be more accurate for students' performance in mathematics computations than for mathematics concept items. Demaray and Elliott (1998) found no difference between correlations in language arts and in mathematics. Hinnant et al. (2009) found that teachers' ratings of academic ability as measured by an academic skills questionnaire were highly correlated with standardized measures of achievement in reading (r = .53–.67) and mathematics (r = .54–.57). Evidently, the empirical findings on the influence of subject matter on teacher judgment accuracy are inconsistent. In addition, there are very few studies focusing on subjects other than language arts and mathematics. As exceptions, achievement in sports (swimming) was measured in a study by Trouilloud, Sarrazin, Martinek, & Guillet, (2002) yielding a comparably high correlation between teachers' judgments and students achievement in swimming (r = .78). Music achievement was measured in a study by Klinedinst (1991), which reported a comparably low correlation (r = .21).

CBM procedures vs. standardized achievement tests. Some studies of teacher judgment accuracy have used CBM procedures as indicators of students' achievement (Eckert et al., 2006; Feinberg & Shapiro, 2003; Hamilton & Shinn, 2003). According to Feinberg and Shapiro (2003), CBM is closely linked to actual in-class student performance, as methods derived from curriculum materials provide a closer overlap with the content of instruction than do published norm-referenced tests. For example, Feinberg and Shapiro (2009) used three reading probes of 150 words to measure students' oral reading fluency. The number of words read correctly per minute was used as a measure of students' reading performance. Mispronunciations, omissions, and substitutions were counted as errors. The median for each of the three probes was computed and used as the student's overall CBM score. The authors argued that teachers are likely to use students' observed classroom behavior as their basis for judging students' academic achievement. They criticized the "lack of a potential overlap between content assessed on a standardized test and student behavior" (Feinberg & Shapiro, 2003; p. 54), which complicates accurate teacher judgments. Indeed, students' achievement on curriculum-based testing material has been compared with their achievement on published, standardized achievement tests. Feinberg and Shapiro (2009) found that correlations between a CBM procedure measuring oral reading fluency and teachers' predictions of oral reading fluency were slightly higher (r = .64) than correlations between a global teacher rating of students' performance and two subtests of a standardized achievement test (r = .59 and r = .60).

Domain specificity. Like teacher judgments, academic achievement tests also differ in terms of their domain specificity. Whereas some tests are designed to measure a very specific academic ability (e.g., phonological awareness; Bailey & Drummond, 2006), others measure different aspects of academic ability (e.g., the Woodcock-Johnson Achievement Battery; Benner & Mistry, 2007).

In terms of test characteristics, we found no evidence for a difference in teacher judgment accuracy between language arts and mathematics in our meta-analysis. The effects of the other test characteristics were not significant either. Therefore, results are generalizable across several types of judgments and tests.

Correspondence Between Judgment and Test Characteristics

As depicted in our model, the correspondence between judgment characteristics and test characteristics is assumed to influence teacher judgment accuracy.

Time gap. In their review, Hoge and Coladarci (1989) included only studies in which the achievement test was administered at the same time as the teacher rating task. In many studies, however, these two measures are not implemented concurrently (Pomplun, 2004). Due to temporal proximity, we expect to find higher correlations between teachers' judgments and students' academic achievement when both measures are administered concurrently than when the test is administered either before or after the rating task.

Congruence in domain specificity. Finally, we considered the congruence in the domain specificity of the teacher rating task and the achievement test. Theoretically, the achievement test may measure a specific academic ability whereas the teacher judgment task may be less specific—or vice versa. For example, Hecht and Greenfield (2001) found teachers' judgments of students' overall academic competence to be correlated with the students' performance on the Letter-Word-Identification subtest of the Woodcock-Johnson Test of Achievement–Revised. Here, a general judgment is set in relation to a very specific ability. We expected to find higher correlations between teachers' judgments and students' achievement in studies in which the domain specificity of the teacher rating task and the achievement test was congruent (e.g., teachers rated students' reading comprehension; students were administered a test of reading comprehension), and lower correlations in studies in which the domain specificity was incongruent (e.g., teachers rated students' overall academic achievement; students were administered a test of reading comprehension).

As expected, the congruence between the teachers' rating task and the achievement test administered to students was related to teacher judgment accuracy, with higher congruence being associated with higher accuracy levels. Because the match between teachers' judgments and students' test performance was higher when both measures addressed the same domain and same ability within a domain, it is reasonable to assume that a "mismatch" leads to lower teacher judgment accuracy.

Correspondence Between Teacher Characteristics and Student Characteristics

To complete the picture of teacher judgment accuracy, our model includes the correspondence between the teacher characteristics and the student characteristics influencing teacher judgment accuracy. To our knowledge, few studies have taken this aspect into account, and their results have been mixed. In their study of teachers' expectations of the reading and mathematics achievement of their students, Alexander et al. (1987) took teacher and student race (Black vs. White) and socioeconomic status into account. The authors found that low-status and Black students were evaluated less favorably than high-status and White students, especially by high-status teachers. Chang and Sue (2003) studied the effects of student race on teachers' assessment of student behavior using vignettes that were paired with a photograph of a child. Due to unequal sample sizes (74.1% European Americans, 10.6% Hispanic, 3.6% African American), the authors were unable to analyze the ratings with respect to teacher ethnicity. In another study, Chang and Demyan (2007) overcame this problem by dichotomizing teachers' race into "White" and "ethnic minority." Teachers rated the students on personal traits; no significant interactions between teacher race and child race were found for any of the 15 traits.

Unfortunately, we were not able to study the effects of the correspondence between teacher characteristics and student characteristics in our meta-analysis as comparable data (e.g., on ethnicity) were not reported in a sufficient number of studies.

EXPERIMENTAL RESEARCH ON TEACHER JUDGMENT ACCURACY

Besides the field studies introduced above, teacher judgment accuracy has also been studied in experimental studies, which allows for inspecting teachers' information processing and decision making more closely (e.g, Krolak-Schwerdt, Böhmer, & Gräsel, 2013; Südkamp & Möller, 2009). In our own empirical research, we have been using a simulated classroom paradigm (Südkamp, Möller, & Pohlmann, 2008).

The Simulated Classroom

The Simulated Classroom (see also Fiedler, Freytag, & Unkelbach, 2007; Fiedler, Walther, Freytag, & Plessner, 2002) is a computer simulation of an instructional situation in which student factors—e.g., achievement (in terms of the proportion of correct answers), motivation (in terms of participation in class), gender (as indicated by a photograph or name)—and instructional factors (subject, number of students, lesson length, content covered) can be experimentally manipulated. The Simulated Classroom is programmed in Java; participants work individually on personal computers. To begin, participants are given an introduction to the functioning of the Simulated Classroom, in which they take on the role of a teacher. Before the lesson starts, they are informed about the students' grade and the topic of the lesson. Their task is to select questions on that topic from a menu of possible questions and to address these questions to the students in their "class." The students are represented by names on virtual desks (see Fig. 2).

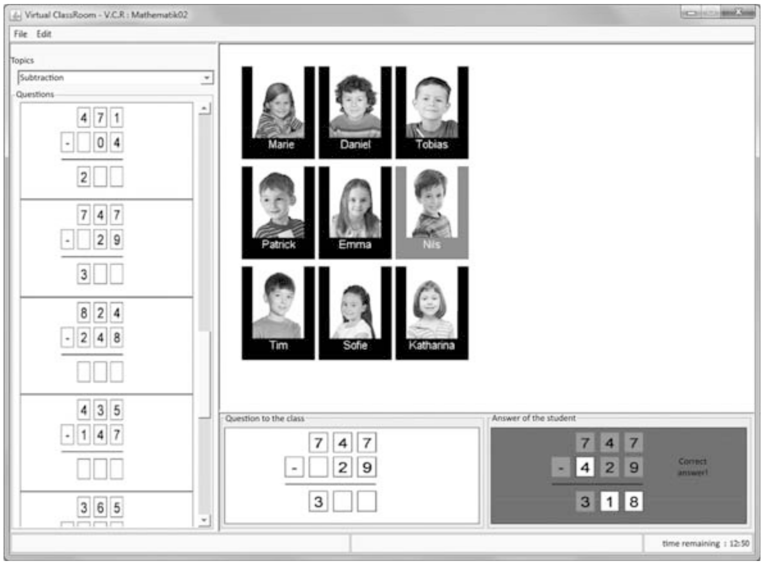

Figure 2. Screenshot of the Simulated Classroom.

The names were chosen at random from the most popular children's names in Germany in the year the simulated student would theoretically have been born. Photographs, names (and thus gender), and seating positions were allocated at random (making sure that the gender and name allocated matched the photo). Each question selected by the "teacher" is displayed at the bottom left of the computer screen. Tasks were either taken from standardized achievement tests or developed specifically according to the study's requirements. In our studies, we mainly focused on the subjects of mathematics and language arts. Bringing together experts in the field of chemistry didactics and educational psychology, Bolte, Köppen, Möller, and Südkamp (2011) developed the Simulated Science Classroom, which incorporates tasks on specific science concepts (e.g. the particle model of matter or concepts of inquiry). When a question has been selected, students volunteer to answer the question in accordance with their predefined motivation parameters. These students are indicated by the coloring of their desks, which changes to yellow. Depending on the study's aim the "teacher" may call on any of the students (whether or not they have volunteered an answer) or just on students volunteering to answer the question, by a mouse click on the respective desk. That student then gives a correct or an incorrect answer depending on his or her predefined ability parameter. The answer is displayed at the bottom right of the screen. If it is correct, it appears in a green box (see Fig. 2). Otherwise, one of several possible incorrect answers appears in a red box. This variation of incorrect answers reduces the probability that the same incorrect answer will be given consecutively by different students. Once a question-and-answer sequence has been completed, the "teacher" can direct either the same question or a new question to any of the students. The length of the lesson can be varied; the teacher can ask any number of questions.

The proportion of correct answers provided by each student is experimentally varied and represents the level of student achievement (e.g., probability of a correct answer approximately .80 for high achievement). The simulated students' achievement behavior in the Simulated Classroom is determined by a probability algorithm, such that the proportion of correct or incorrect answers given by each student corresponded approximately with his or her achievement parameter. In most of our studies, we also varied student motivation experimentally, which is operationalized as the probability of a simulated student volunteering to answer a question (e.g., probability of volunteering .20 for low motivation). Motivation behavior in the Simulated Classroom is again determined by a probability algorithm, such that proportion of questions each student volunteered to answer corresponded approximately with his or her motivation parameter. Participants usually have 16-18 minutes to get a picture of the simulated students' academic achievement and motivation.

At the end of the "lesson," participants are asked to judge the proportion of correct answers given by each student. Ratings are usually given on a scale from 0 to 100%. It is thus possible to gauge the extent to which the participants' judgments correspond with the students' actual achievement. Likewise, the proportion of questions that each student volunteers to answer serves as a measure of student motivation. Participants are also asked to judge this proportion, thus allowing perceived and actual student motivation to be compared. Ratings are again given on a percentage scale. To get an indirect judgment, teachers are also asked to grade the simulated students.

Teacher Judgment Accuracy within the Simulated Classroom

Our findings on the accuracy of judgments of student achievement indicate that teachers and teacher candidates are fairly successful in gauging individual students' relative achievement level in the Simulated Classroom. The correlations between actual student achievement and judgments of achievement are usually moderate to high in size. For example, Südkamp et al. (2008) found correlations of .62 (first simulated lesson) and .68 (second simulated lesson), while Kaiser, Retelsdorf, Südkamp, and Möller (2013) found correlations of .69 (second partial study) and .57 (third partial study). These finding are consistent with the results of field studies on the accuracy of teacher judgments of student achievement (Hoge & Coladarci, 1989; Südkamp et al., 2012).

Moderators of Teacher Judgment Accuracy

Concerning moderators of teachers' judgment accuracy, we will focus on the influence of student characteristics and teacher characteristics here, as those factors could not be included in our analyses within the meta-analysis of field studies.

Student characteristics: Motivation and ethnicity. In a recent study based on the simulated classroom paradigm (Kaiser et al., 2013), we combined the field and experimental approach to the study of teacher judgment accuracy and examined whether students' achievement influences teachers' judgments of their motivation and vice versa. In the field study teachers and their students, who were tested and surveyed within a comprehensive project on the development of reading literacy and reading motivation at secondary level in Germany (e.g., Möller, Retelsdorf, Köller, & Marsh, 2011; Retelsdorf, Becker, Köller, & Möller, 2011; Retelsdorf, Köller, & Möller, 2011). Measures of students' academic achievement and motivation were implemented, while teachers' judged students' academic achievement and motivation respectively. In two experimental studies, teacher candidates worked on the Simulated Classroom. Here, the academic achievement and motivation of simulated students was varied experimentally. In all three studies, structural equation modeling revealed an effect of student achievement on teachers' judgments of student motivation and an effect of student motivation on teacher judgments of student achievement—above and beyond the association of each student characteristic with teachers' judgments of that characteristic.

In two studies on the influence of students' ethnicity on teachers' judgments of students' academic achievement, we varied students' ethnicity within the Simulated Classroom (Kaiser, Schubert, Südkamp, & Möller, 2012). In the first study, there were eight German students, one Asian student, and three Turkish students within the Simulated Classroom, which was indicated by their pictures and traditional names. In Germany, having a Turkish migration background is still associated with lower academic achievement in comparison to non-migrant students and with lower academic opportunities. According to this stereotype, we expected Turkish students to be judged less positively compared to their German classmates, while controlling for the students' achievement within the Simulated Classroom. Results showed that low achieving Turkish students were judged less favorably than low achieving German students, but also more accurately. In contrast, high achieving Turkish students were judged more positively than high achieving German students, meaning Turkish students were judged more accurately.

The second study was conducted analogous to the first study. Here, there were eight German students, one Turkish student and three Asian students in the Simulated Classroom. In western society, Asians are often perceived to be a "model minority" showing high effort and achievement in academics. According to this stereotype, we expected the Asian students within the classroom to be judged more positively than their German classmates. Results showed that low achieving Asian students were judged less favorably than low achieving German students, but also more accurately.

On the other hand, high achieving Asian students were judged more positively than high achieving German students, again meaning Asian students were judged more accurately. As we didn't find the expected general positive bias in judgments towards Asian students, we also analyzed whether it could be that being a minority is to account for higher judgment accuracy. So we estimated a path model with manifest variables using the Mplus software (Muthén & Muthén, 2010). We analyzed whether being a minority moderated the relationship between students' actual achievement and teachers' judgment of student achievement. The results revealed a significant path from students' achievement to teachers' judgments, no significant path from the students' minority status, but a significant interaction-path. The significant path from students' actual achievement to teachers' judgments shows, that teachers were quite accurate in judging students' achievement. But the significant interaction reveals that the relationship between students' actual achievement and teachers' judgment is more accurate for minority students.

Teacher characteristics: Cognitive abilities. As mentioned above, there are only a few studies, which explicitly focus on teacher characteristics as moderators of teacher judgment accuracy. To our knowledge, teachers' cognitive abilities as a prerequisite of judgment accuracy have not been a focus of research at all. Since teachers are confronted with much diagnostic information during their lessons, a high information processing speed could be a necessary condition to be able to judge students' characteristics accurately. The aim of this study (see Kaiser, Helm, Retelsdorf, Südkamp, & Möller, 2012 for a more detailed description) was to give a first hint on the relationship between teachers' diagnostic and cognitive abilities. The diagnostic ability of teacher students at the University of Kiel was tested within the Simulated Classroom paradigm. Preceding the Simulated Classroom the subjects' cognitive abilities were measured with 34 selected items of the Advanced Progressive Matrices (APM; Raven, 1962) for one sample. In another sample comprising teacher students the cognitive abilities were measured with the subscale Figure Analogies of the German version of the Cognitive Abilities Test (Heller & Perleth, 2000). In order to test the relation between cognitive and diagnostic abilities a multi level-analysis was conducted. On level 1 students' actual achievement was used to predict teachers' judgments (as a measure of judgment accuracy). On level 2 the cognitive abilities were taken into account and it was tested whether there is a relation between cognitive abilities and judgment accuracy. The multi level-analysis revealed a significant cross-level-interaction for both samples showing that teachers' higher cognitive abilities were associated with higher judgment accuracy.

SUMMARY AND OUTLOOK

In this book chapter, we first extended the description of our heuristic model of teacher judgment accuracy, which was briefly introduced in the discussion of Südkamp et al. (2012). In line with the model, we brought together findings on teacher judgment accuracy from field studies. As empirical research findings of 75 studies on the issue are statistically summarized in our meta-analysis, we highlight the most important findings here. Second, we introduced an experimental approach to the study of teacher judgment accuracy. In our own empirical research, the Simulated Classroom proved to be a useful instrument for examining teacher judgment accuracy. The instrument makes it possible to experimentally manipulate student characteristics, and thus complements studies conducted in real-life contexts: Psychological phenomena that have been observed in field studies to be associated with teacher judgments of students can be investigated more closely under experimental conditions (Wang, Treat, & Brownell, 2008).

Although our findings concerning teacher judgment accuracy are similar in field and experimental studies, it remains questionable to what extent the same findings can be expected in real-life classrooms as in the Simulated Classroom. In the following, we discuss some differences between teacher judgments in the two contexts. Whereas in real life, teachers' judgments are based on a wealth of informal observations, in the Simulated Classroom, it is possible to experimentally control the information on which these judgments are based. The observations made in the Simulated Classroom are thus far less complex than those made in real-life classroom situations, in which the demands on teachers in terms of the collection and interpretation of diagnostic information are much higher. The reduced complexity of the Simulated Classroom may thus be a further explanation for the slightly higher levels of judgment accuracy observed in the experimental studies. Moreover, the systematic distribution of student achievement and motivation implemented in the Simulated Classroom does not reflect that found in a natural environment. The instrument can, however, be extended to allow the targeted variation of various student and class characteristics, thus allowing a differentiated analysis of information processing processes in teachers and teacher candidates. The Simulated Classroom is therefore considered a promising tool for the research concerning teachers' judgment accuracy (Schrader, 2010; Spinath, 2012).

The research presented in this book chapter brings together different research strategies. However, the question whether the same construct is measured with the different approaches still needs to be clarified empirically. This will be the focus of future projects. As a long-term objective the Simulated Classroom could be used as a training instrument. Teachers and teacher candidates could get immediate feedback regarding their judgment accuracy and as a result decrease erroneous judgment tendencies.

REFERENCES/BIBLIOGRAPHY

Alexander, K. L., Entwisle, D. R., & Thompson, M. S. (1987). School performance, status relations, and the structure of sentiment: Bringing the teacher back in. *American Sociological Review, 52*, 665–682. doi:10.2307/2095602

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

Bailey, A. L., & Drummond, K. V. (2006). Who is at risk and why? Teachers' reasons for concern and their understanding and assessment of early literacy. *Educational Assessment, 11*, 149–178. doi:10.1207/s15326977ea1103&4_2

Baron, R. M., Tom, D. Y. H., & Cooper, H. M. (1985). Social class, race, and teacher expectations. In J. B. Dusek, V. C. Hall, & W. J. Meyer (Eds.), *Teacher expectancies*. Hillsdale, NJ: Erlbaum.

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

Begeny, J. C., Eckert, T. L., Montarello, S. A., & Storie, M. S. (2008). Teachers' perceptions of students' reading abilities: An examination of the relationship between teachers' judgments and students' performance across a continuum of rating methods. *School Psychology Quarterly, 23*, 43–55. doi:10.1037/1045-3830.23.1.43

Benner, A. D., & Mistry, R. S. (2007). Congruence of mother and teacher educational expectations and low-income youth's academic competence. *Journal of Educational Psychology, 99*, 140–153. doi:10.1037/0022-0663.99.1.140

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*, 347–356. doi:10.1037/0022-0663.85.2.347

Bolte, C., Köppen, G., Möller, J., & Südkamp, A. (2012). Diagnostic competence of pre-service teachers analysed by means of the simulated science classroom. In C. Bruguière, A. Tiberghien, & P. Clément (Eds.), *E-Book Proceedings of the ESERA 2011 Conference: Science learning and Citizenship*. Part 13 (J. Viiri & D. Couso, pp. 18–24). Lyon, France: European Science Education Research Association.

Chang, D. F., & Demyan, A. L. (2007). Teachers' stereotypes of Asian, Black, and White students. *School Psychology Quarterly, 22*, 91–114. doi:10.1037/1045-3830.22.2.91

Chang, D. F., & Sue, S. (2003). The effects of race and problem type on teachers' assessments of student behavior. *Journal of Consulting and Clinical Psychology, 71*, 235–242. doi:10.1037/0022-006X.71.2.235

Coladarci, T. (1986). Accuracy of teacher judgments of student responses to standardized test items. *Journal of Educational Psychology, 78*, 141–146. doi:10.1037/0022-0663.78.2.141

Demaray, M. K., & Elliott, S. N. (1998). Teachers' judgments of students' academic functioning: A comparison of actual and predicted performances. *School Psychology Quarterly, 13*, 8–24. doi:10.1037/h0088969

DuPaul, G. J., Rapport, M. D., & Perriello, L. M. (1991). Teacher ratings of academic skills: The development of the Academic Performance Rating Scale. *School Psychology Review, 20*, 284–300.

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

Eckert, T. L., Dunn, E. K., Codding, R. S., Begeny, J. C., & Kleinmann, A. E. (2006). Assessment of mathematics and reading performance: An examination of the correspondence between direct assessment of student performance and teacher report. *Psychology in the Schools, 43*, 247–265. doi:10.1002/pits.20147

Feinberg, A. B., & Shapiro, E. S. (2003). Accuracy of teacher judgments in predicting oral reading fluency. *School Psychology Quarterly, 18*, 52–65. doi:10.1521/scpq.18.1.52.20876

Feinberg, A. B., & Shapiro, E. S. (2009). Teacher accuracy: An examination of teacher-based judgments of students' reading with differing achievement levels. *The Journal of Educational Research, 102*, 453–462. doi:10.3200/JOER.102.6.453-462

Fiedler, K., Freytag, P., & Unkelbach, C. (2007). Pseudocontingencies in a simulated classroom. *Journal of Personality and Social Psychology, 92*, 665–677. doi:10.1037/0022-3514.92.4.665

Fiedler, K., Walther, E., Freytag, P., & Plessner, H. (2002). Judgment biases in a simulated classroom: A cognitive–environmental approach. *Organizational Behavior and Human Decision Processes, 88*, 527–561. doi:10.1006/obhd.2001.2981

Fletcher, J., Tannock, R., & Bishop, D. V. M. (2001). Utility of brief teacher rating scales to identify children with educational problems: Experience with an Australian sample. *Australian Journal of Psychology, 53*, 63–71. doi:10.1080/00049530108255125

Graney, S. B. (2008). General education teacher judgments of their low-performing students' short-term reading progress. *Psychology in the Schools, 45*, 537–549. doi:10.1002/pits.20322

Hamilton, C., & Shinn, M. R. (2003). Characteristics of word callers: An investigation of the accuracy of teachers' judgments of reading comprehension and oral reading skills. *School Psychology Review, 32*, 228–241.

Hauser-Cram, P., Sirin, S. R., & Stipek, D. (2003). When teachers' and parents' values differ: Teachers' ratings of academic competence in children from low-income families. *Journal of Educational Psychology, 95*, 813–820. doi:10.1037/0022-0663.95.4.813

Hecht, S. A., & Greenfield, D. B. (2001). Comparing the predictive validity of first grade teacher ratings and reading-related tests on third grade levels of reading skills in young children exposed to poverty. *School Psychology Review, 30*, 50–69.

Heller, K. A., & Perleth, C. (2000). *Kognitiver Fähigkeitstest für 4. bis 12. Klassen, Revision. [Cognitive Abilities Test (CogAT; Thorndike, L. & Hagen, E., 1954-1986)--German adapted version/author)].* Göttingen: Beltz.

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

Helwig, R., Anderson, L., & Tindal, G. (2001). Influence of elementary student gender on teachers' perceptions of mathematics achievement. *Journal of Educational Research*, *95*, 93–102. doi:10.1080/00220670109596577

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

Hoge, R. D., & Butcher, R. (1984). Analysis of teacher judgments of pupil achievement levels. *Journal of Educational Psychology, 76*, 777–781. doi:10.1037/0022-0663.76.5.777

Hoge, R. D., & Coladarci, T. (1989). Teacher-based judgments of academic achievement: A review of literature. *Review of Educational Research, 59*, 297–313. doi:10.2307/1170184

Hopkins, K. D., George, C. A., & Williams, D. D. (1985). The concurrent validity of standardized achievement tests by content area using teachers' ratings as criteria. *Journal of Educational Measurement, 22*, 177–182. doi:10.1111/j.1745-3984.1985.tb01056.x

Hurwitz, J. T., Elliott, S. N., & Braden, J. P. (2007). The influence of test familiarity and student disability status upon teachers' judgments of students' test performance. *School Psychology Quarterly, 22*, 115–144. doi:10.1037/1045-3830.22.2.115

Impara, J. C., & Plake, B. S. (1998). Teachers' ability to estimate item difficulty: A test of assumptions in the Angoff standard setting method. *Journal of Educational Measurement, 35*, 69–81. doi:10.1111/j.1745-3984.1998.tb00528.x

Jussim, L. J., & Eccles, J. (1995). Are teacher expectations biased by students' gender, social class, or ethnicity? In Y.-T. Lee, L. J. Jussim, & C. R. McCauley (Eds.), *Stereotype accuracy: Toward appreciating group differences* (pp. 245–271). Washington, DC: American Psychological Association.

Jussim, L., Eccles, J. S., & Madon, S. (1996). Social perception, social stereotypes, and teacher expectations: Accuracy and the quest for the powerful self-fulfilling prophecy. *Advances in Experimental Social Psychology, 28*, 281–388. doi:10.1016/S0065-2601(08)60240-3

Kaiser, J., Helm, F., Retelsdorf, J., Südkamp, A., & Möller, J. (2012). Zum Zusammenhang von Intelligenz und Urteilsgenauigkeit bei der Beurteilung von Schülerleistungen im Simulierten Klassenraum [On the relation of intelligence and judgment accuracy in the process of assessing student achievement in the simulated classroom]. *Zeitschrift für Pädagogische Psychologie/German Journal of Educational Psychology, 26*, 251–261. doi:10.1024/1010-0652/a000076

Kaiser, J., Retelsdorf, J., Südkamp, A., & Möller, J. (2013). *Achievement and engagement: How student characteristics influence teacher judgments.* Manuscript submitted for publication.

Kaiser, J., Schubert, C., Südkamp, A., & Möller, J. (2012, September). *Stehen Minderheiten im Fokus bei der Beurteilung durch Lehrkräfte? [Do teachers focus minorities when making judgments?].* Paper presented at the 48th Congress of the German Psychological Society (DGPs), Bielefeld, Germany.

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/ German Journal of Educational Psychology, 23,* 197–209. doi:10.1024/1010-0652.23.34.197

Kennedy, E. (1995). Contextual effects on academic norms among elementary school students. *Educational Research Quarterly, 18*, 5–13.

Kenny, D. T., & Chekaluk, E. (1993). Early reading performance: A comparison of teacher-based and test-based assessments. *Journal of Learning Disabilities, 26*, 227–236. doi:10.1177/002221949302600403

Klinedinst, R. E. (1991). Predicting performance achievement and retention of fifth-grade instrumental students*. Journal of Research in Music Education, 39*, 225–238. doi:10.2307/3344722

Krolak-Schwerdt, S., Böhmer, M., & Gräsel, C. (2013). The impact of accountability on teachers' assessments of student performance: A social cognitive analysis. *Social Psychology of Education*. Advance online publication. doi:10.1007/s11218-013-9215-9

Kuklinski, M. R., & Weinstein, R. S. (2001). Classroom and developmental differences in a path model of teacher expectancy effects. *Child Development, 72*, 1554–1578. doi:10.1111/1467-8624.00365

Li, H., Pfeiffer, S. I., Petscher, Y., Kumtepe, A. T., & Mo, G. (2008). Validation of the Gifted Rating Scales–School Form in China. *Gifted Child Quarterly, 52*, 160–169. doi:10.1177/0016986208315802

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

Maguin, E., & Loeber, R. (1996). How well do ratings of academic performance by mothers and their sons correspond to grades, achievement test scores, and teachers' ratings? *Journal of Behavioral Education, 6*, 405–425. doi:10.1007/BF02110514

Möller, J., Retelsdorf, J., Köller, O., & Marsh, H. W. (2011). The Reciprocal I/E Model: An integration of models of relations between academic achievement and self-concept. *American Educational Research Journal, 48*, 1315–1346. doi:10.3102/0002831211419649

Mulholland, L. A., & Berliner, D. C. (1992, April). *Teacher experience and the estimation of student achievement.* Paper presented at the Annual Meeting of the American Educational Research Association, San Francisco, CA.

Muthén, L. K., & Muthén, B. O. (2010). *Mplus (Version 6)* [Computer software]. Los Angeles, CA: Muthén & Muthén.

Pomplun, M. (2004). The differential predictive validity of the initial skills analysis: Reading screening tests for K-3. *Educational and Psychological Measurement, 64*, 813–827. doi:10.1177/0013164404263879

Raven, J. C. (1962). *Advanced progressive matrices*. London: Lewis & Co. Ltd.

Retelsdorf, J., Becker, M., Köller, O., & Möller, J. (2011). Reading development in a tracked school system: A longitudinal study over 3 years using propensity score matching. *British Journal of Educational Psychology, 82*, 647–671. doi:10.1111/j.2044-8279.2011.02051.x

Retelsdorf, J., Köller, O., & Möller, J. (2011). On the effects of motivation on reading performance growth in secondary school. *Learning and Instruction, 21*, 550–559. doi:10.1016/j.learninstruc.2010.11.001

Ritts, V., Patterson, M. L., & Tubbs, M. E. (1992). Expectations, impressions and judgments of physically attractive students: A review. *Review of Educational Research, 62*, 413–426. doi:10.3102/00346543062004413

Schrader, F.-W. (2010). Diagnostische Kompetenz von Eltern und Lehrern. [Diagnostic competence of parents and teachers.] In D. H. Rost (Ed.), *Handwörterbuch Pädagogische Psychologie* (4th ed., pp. 102–108). Weinheim: Beltz.

Schrader, F.-W., & Helmke, A. (2001). Alltägliche Leistungsbeurteilung durch Lehrer [Day-to-day performance evaluation by teachers]. In F. E. Weinert (Ed.), *Leistungsmessung in Schulen* (pp. 45–58). Weinheim, Germany: Beltz.

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

Spinath, B. (2012). Beiträge der Pädagogischen Psychologie zur Professionalisierung von Lehrerinnen und Lehrern: Diskussion zum Themenschwerpunkt. [Educational Psychology's contributions to professional teacher development: Discussion of the special issue]. *Zeitschrift für Pädagogische Psychologie/German Journal of Educational Psychology, 26*(4), 307–312. doi:10.1024/1010-0652/a000082

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

Tenenbaum, H. R., & Ruck, M. D. (2007). Are teachers' expectations different for racial minority than for European American students? A meta-analysis. *Journal of Educational Psychology, 99*, 253–273. doi:10.1037/0022-0663.99.2.253

Tiedemann, J. (2000). Parents' gender stereotypes and teachers' beliefs as predictors of children's concept of their mathematical ability in elementary school. *Journal of Educational Psychology, 92*, 144–151. doi:10.1037/0022-0663.92.1.144

Trouilloud, D. O., Sarrazin, P. G., Martinek, T. J., & Guillet, E. (2002). The influence of teacher expectations on student achievement in physical education classes: Pygmalion revisited. *European Journal of Social Psychology, 32*, 591–607. doi:10.1002/ejsp.109

Wang, S. S., Treat, T. A., & Brownell, K. D. (2008). Cognitive processing about classroom-relevant contexts: Teachers' attention to and utilization of girls' body size, ethnicity, attractiveness, and facial affect. *Journal of Educational Psychology, 100*, 473–489. doi:10.1037/0022-0663.100.2.473

Wigfield, A., Galper, A., Denton, K., & Seefeldt, C. (1999). Teachers' beliefs about former Head Start and non-Head Start first-grade children's motivation, performance, and future educational prospects. *Journal of Educational Psychology, 91*, 98–104. doi:10.1037/0022-0663.91.1.98

AFFILIATIONS

Anna Südkamp
Rehabiliation Psychology
TU Dortmund University, Germany

Johanna Kaiser
Educational Psychology
University of Kiel, Germany

Jens Möller
Educational Psychology
University of Kiel, Germany

CORDULA ARTELT & TOBIAS RAUSCH

ACCURACY OF TEACHER JUDGMENTS

When and for What Reasons?

INTRODUCTION

Assessing students' competencies (either implicitly or explicitly) is a routine task for teachers, with many decisions – for example, concerning grading, ability grouping, and difficulty level of questions or materials – being based on these teacher judgments. Given this, judgment accuracy (often referred to as diagnostic competence) is regarded as a central element of teacher professionalism and is thought to affect students' learning outcomes. So far, however, empirical evidence concerning teachers' diagnostic competence is incomplete and partly contradictory. As argued throughout this paper, these mixed results can be attributed, to some extent, to the fact that judgment purposes and judgment demands are not adequately taken into account in current research literature.

ACCUARY OF TEACHER JUDGMENTS

The Concept of Diagnostic Competence[1]

The term diagnostic competence refers to the ability to judge a person's characteristics and attributes correctly. Accordingly, teachers' diagnostic competence refers to their ability to judge student characteristics (e.g., student achievement) correctly (Schrader, 2010). The term also refers to teachers' ability to correctly judge task demands (Anders, Kunter, Brunner, Krauss, & Baumert, 2010). It is regarded as a key competence in the context of teaching and learning, since it is a prerequisite for adequate classroom organization and adaptive teaching, as well as being the basis of pedagogical decisions and actions (task selection, learning feedback and evaluation).

In recent years, however, there has been quite a debate about the concept of diagnostic competence. One of the reasons why the term diagnostic competence has been criticized is the lack of empirical support for the assumption of a homogeneous competence dimension. It appears to be the case that teachers' judgment accuracy varies as a function of the respective achievement/competence domain under study (e.g., students' mathematical or reading competence, their motivation, anxiety, or intelligence; e.g., Spinath, 2005), as a function of the specific accuracy indicator (e.g., the level-, rank order- or variation-component of teachers' judgments; e.g., Karst, 2012), as well as a function of the concrete judgment demands (e.g., Südkamp, Kaiser

S. Krolak-Schwerdt et al., (Eds.), Teachers' Professional Development, 27–43.

& Möller, 2012). To this end, there is reason to believe that diagnostic competence is not an adequate term for these intra-individually varying judgment qualities. On the other hand, broader models of diagnostic competence (e.g., van Ophuysen, 2010) claim that diagnostic competence should not be reduced to accuracy measures, but rather should be seen as pedagogical decision making in everyday professional school practice. Judgment accuracy is thereby regarded as one facet of the outcome of diagnostic competence. In addition, it also includes other aspects of the decision making process, namely the quality of the information basis, as well as the quality of the process of pedagogical-diagnostic decision making. Accordingly, Schrader (2011) argues that a broader understanding of the concept of diagnostic competence concept should cover the entire process of diagnostic decision making, starting with the collection of data and right up to the evaluation of an intervention which might then be conducted on the basis of the derived diagnostic decisions. To this end, judgment accuracy is only a small part and not the competence per se.

Apart from the discussion about the validity of the term diagnostic competence, there is a broad consensus that a variety of teacher activities are heavily dependent on teachers' ability to correctly assess students' competencies and the demands of tasks and didactic units. Many professional decisions are based on teacher judgments of student characteristics; for example, decisions related to ability grouping, adaptive teaching, grade allocation, as well as deciding on task difficulty levels, and creating tests or assessments in classroom. A more indirect example of the relevance of teacher judgment is the specific kind of causal attribution teachers use when evaluating student success or failure. Teachers who incorrectly interpret a student's failure as a lack of ability (internal / stable) or effort (internal / variable) are not only likely to be biased when allocating grades, but also behave differently in terms of further achievement expectations, which is likely to impact students' self-concept. Consequently, judgment accuracy is regarded as a central element of teacher professionalism and is assumed to affect students' learning outcomes. Whether it is reasonable to use the term diagnostic competence (as a trait or disposition) or to use the term judgment accuracy instead (as state) is still a matter of debate.

Indicators of Judgment Accuracy

The quality of teacher judgments can be estimated using different indicators and judgment types. A common classification of indicators differentiates between the accuracy of teacher judgments in terms of the level of student performance, the variation of student performance, and the accuracy of the rank order of student performance (e.g., Schrader, 2010). According to Helmke, Hosenfeld, and Schrader (2004), the rank-order component can be regarded as the central element of teacher judgment accuracy. When calculating this indicator, teachers' judgments about the rank order of their students with respect to a specific characteristic (e.g., achievement or competence) are correlated with the rank order of students based on standardized tests assessing the particular student characteristic. Accordingly, class-specific correlations are computed, resulting in only one indicator per teacher and class. The majority of studies indicate that teachers' rank-order judgments correlate fairly highly with the rank order of their students: based on their 1989 meta-analysis, Hoge and Coladarci reported correlations between $r = .28$ and $r = .92$, with a median correlation of $r = .66$. In a recent meta-analysis (Südkamp et al., 2012), the Fisher's z-transformed correlations ranged between $r = -.03$ and $r = 1.18$, with a median correlation of $r = .53$. However, as can be seen by the range of results of the different studies reported in the two meta-analyses, there are considerable differences between studies. Moreover, the range of individual scores within each study is also considerable. Teachers seem to vary in the level of accuracy of their rank-oder judgment.

Less research has been conducted focusing on two other aspects of teacher judgments: the level component and the variation component. However, with respect to the level component, it seems that teachers as well as teacher candidates tend to overestimate students' achievement level (e.g., Bates & Nettelbeck, 2001; Helmke et al., 2004; Südkamp & Möller, 2009). The results of the variation component are less clear. While there are studies showing that teachers underestimate the variance of class achievement (e.g., Helmke et al., 2004; Südkamp & Möller, 2009), the literature also shows evidence of overestimations, as well as instances of accurate judgments (Brunner, Anders, Hachfeld, & Krauss, 2011).

The three indicators of judgment accuracy can be estimated by using different types of teacher judgments, as well as assessments of the corresponding student characteristics. The most common approach is a judgment type in which the teachers are asked to judge students' academic achievement on a rating scale (e.g., ranging from poor to excellent). Teachers' ratings are then correlated with students' test performance on a corresponding achievement test. All three indicators can be calculated for this kind of (global) rating; however, in order to be able to interpret the level- and the variation component, further more or less arbitrary assumptions (i.e., concerning the reference point for accurate judgments) as well as transformations of the original scales are needed. Most research using global judgments (ratings) thus focuses on the rank-order component as the indicator of judgment accuracy. Other, more specific, indicators use teacher judgments of individual students' performance on particular items and estimate judgment accuracy in terms of the rank order, the variation, and the level component by mapping the teacher ratings to the actual performance of the student on the specific tasks. Thus, judgment types clearly differ with respect to their specificity. Consequently, judgment specificity has been used as a moderator variable in meta-analyses (see below). Furthermore, the task-specific judgments can be used to build additional indicators. An indicator that uses all of the available information of the judgment of individual students' performance on individual tasks is the task-specific hit rate (Karing, Matthäi, & Artelt, 2011). Thus, a specific feature of the task-specific hit rate is the fact that it takes into account whether the tasks that were rated by the teachers as having been solved by the students are the same tasks that were indeed solved by the students. With less specific judgments, this would not be possible, as becomes obvious in the following example: a teacher estimates the number of correct items in his or her class by reporting the percentage of correct answers per class. This percentage might be the same as the one estimated by aggregating individual student responses on these items. However, teachers might have a different perception of the individual difficulty of the items, leading to a perfect correspondence for the class level judgment, but a low correspondence for the (aggregated) indicator of the task-specific hit rate. The task-specific hit rate thus covers more aspects of the diagnostic decision-making task. In addition to the task-specific hit rate, the amount of over- or underestimation (non-hit rate) can be analysed by considering the number of incorrect hits (see Karing, Matthäi, & Artelt, 2011)[2]. The hit rate indicators allow for detailed analyses of differential judgment accuracy – for example, dependent on task difficulty, task demands or type of task – which is important information for teacher training and other attempts at fostering teacher judgment accuracy. However, at this stage, there are only a few studies that use the exact mapping of teacher judgments and student performance based on task-specific judgments. Demaray and Elliott (1998) used such an indicator ("performance/judgment agreement" in their terminology) and found moderately high levels of judgment accuracy. In an even older study, Coladarci (1986) found an average of 73 percent correct judgments of primary school teachers' task-specific judgments of students' reading comprehension test performances. In our own studies based on data of the BiKS Study in Bamberg[3] (see Table 1), the task-specific hit rate – for teachers of secondary schools (grade 5, 6, 7, and 8) – were lower, ranging between 53 percent to 66 percent for reading comprehension, and 60 percent to 66 percent for mathematics.

Taken together, the two new indicators (i.e., task-specific hit rate and non-hit rate) complement the three indicators (rank-order, level and variation component), since they more specifically take into account to what extent teacher judgments and student performances correspond to each other on the item level. Task-specific indicators, however, can only be computed when teachers estimate student performance on specific tasks. As will be argued throughout this chapter, it matters for the quality of teacher judgments which kind of judgment (global vs. task-specific) teachers are asked to perform, and thus which indicators are used to estimate teacher accuracy, as well as its effects.

Potential Moderators of Judgment Accuracy

When looking at inter-individual differences in teacher judgment accuracy, features of the judgment task, as well as characteristics of teachers and students involved in the judgment process, need to be taken into account. Südkamp and colleagues (2012) introduced a heuristic model of teacher based judgments of students' academic achievement that differentiates moderator variables according to four major categories: student characteristics, teacher characteristics, judgment characteristics, and test characteristics. Their meta-analysis focuses primarily on moderators of the latter two categories: judgment and test characteristics (as well as their correspondence). This is partly because the research on teacher characteristics has not yet provided evidence for individual teacher characteristics that influence judgment accuracy, while research on student characteristics produces mixed results.

Judgment characteristics. In their meta-analysis, Hoge and Coladarci (1989) distinguished – among others – between direct and indirect assessments. Südkamp and colleagues (2012) use the term 'informed versus uninformed ratings' instead, arguing that the lack of a transparent standard of comparison (i.e., information about the student assessment test) is the main feature that characterizes indirect, as opposed to direct, judgments according to Hoge and Coladarci (1989). In both meta-analyses the specificity of teachers' ratings were studied as a potential moderator variable. Südkamp and colleagues (2012) classified judgment specificity along the category ratings (e.g., rating of students' performance in mathematics), rankings (e.g., ranking students from lowest to highest in reading ability) or estimations of correct responses (e.g., estimation of solved items). In addition to informed versus uninformed judgments and judgment specificity, Südkamp and colleagues (2012) took the following features into account: number of points on the rating scale, norm-referenced versus peer-dependent judgments, and the domain specificity of teachers' judgments. Additionally, as moderators at the test level, subject matter and curriculum based measures versus standardized achievement tests were tested. Finally, they took into account the time gap between test and judgment, as well as the congruence between test and judgment in domain specificity. Among the potential moderator variables, only two were empirically proven: teachers' judgment accuracy was moderated by use of either informed or uninformed teacher judgments, with use of informed judgments leading to a higher correspondence between teachers' judgments and students' academic achievement. Hoge and Coladarci's (1989) meta-analysis came to the same conclusion, reporting largely the same phenomena and results. The second significant predictor reported by Südkamp and colleagues (2012) was congruence in domain specificity, meaning that lower accuracy estimates were found in studies in which the domain specificity was incongruent (e.g., teachers rated students' overall academic achievement, whereas students were administered a test of reading comprehension). Judgment specificity did not seem to make a difference. However, Hoge and Coladarci (1989) reported differences for specific versus global judgments: teachers' global judgments (ratings) correlated a little lower with students' test performance than was the case for task-specific judgments (r = .61 vs. r = .70). However, it has to be taken into account that the number of studies that actually used task-specific judgments is very limited, and that these studies partly exhibit other characteristics that might be relevant in this respect. For example, Feinberg and Shapiro (2009) studied the task-specific judgments of teachers who were given additional and very detailed information about the test situation, possibly leading to higher levels of accuracy. Demaray and Elliott (1998), on the other hand, used a global scale that included a rating on students' reading comprehension, as well as their motivation and general academic performance (low and high congruence in domain specificity according to Südkamp et al., 2012) which likely caused lower scores for the global ratings.

Accuracy of Global Versus Task-Specific Judgments

As mentioned above, our own research project which is part of the BiKS research group in Bamberg examines research questions on prerequisites, structure, and effects of teachers' diagnostic competence by implementing different assessments for teachers and their students. Findings of the research project are reported throughout the whole chapter. In the remainder of this chapter, we specifically compare results using teachers' global ratings of students' competence with results using teachers' task-specific judgments of students' test performance. Thus, we use teacher judgments about students' competence based on global ability concepts and without reference to a specific test (uninformed judgments according to Südkamp et al., and indirect and unspecific judgments according to Hoge and Coladarci). These global judgments are teachers' estimations about general ability level in a particular domain (e.g., judging students' abilities in mathematics along a 5-point rating scale, compared to an average student of the same age). At the same time, we use teachers' task-specific judgments of individual students' performance on a number of particular tasks (informed and specific judgments according to Südkamp et al., and direct and specific according to Hoge and Coladarci). The judgments are therefore either located on a global level or on a task-specific level. We estimated the rank-order indicator for global judgments and the rank-order indicator as well as the task-specific hit rate for the task-specific ratings for teachers of grade 5 to 8. The results are depicted in Table 1.

Table 1. Judgment accuracy for reading and mathematics according to different indicators

Mean/SD (N)	*Grade 5*	*Grade 6*	*Grade 7*	*Grade 8*
Reading comprehension				
Rank-order - global	.34/.33 (94)	.37/.37 (82)	.36/.40 (74)	.64/.26 (15)
Rank-order - task-specific	.19/.33 (66)	.19/.35 (63)	.20/.43 (63)	.33/.39 (13)
Hit rate - task-specific*	59/7 (66)	62/10 (63)	66/10 (63)	53/8 (14)
Mathematical competence				
Rank-order - global	.57/.38 (73)	.56/.35 (79)	.53/.46 (79)	.66/.59 (29)
Rank-order - task-specific	.46/.41 (72)	.40/.35 (77)	.41/.42 (74)	.44/.40 (24)
Hit rate - task-specific*	63/12 (72)	66/10 (77)	60/10 (74)	61/9 (24)

Note: * in percent

For the mathematical domain, the results for the rank-order component (global judgments) are roughly the same as those reported in the two meta-analyses. The results for the rank-order components based on task-specific judgments, however, tend to be lower. This is also the case for the domain of reading, although both types of rank-order indicators produce lower results for reading than for mathematics. The third indicator (hit rate, based on task-specific teacher judgments) shows the percentage of correct teacher judgments about students' performance on a number of items per domain. Overall, the hit rate is not very high, with just over 50% of the items rated correctly. It seems to be the case that global judgments are easier – in the sense that they produce more accurate results – than task-specific judgments (see also Karing, Matthäi, & Artelt, 2011).

WHEN CAN WE EXPECT TO FIND ACCURATE JUGMENTS?

Obviously, there are accuracy differences in teachers' judgments, both within and between studies. In order to better understand this variation, it is worth having a closer look at the aspects of the situations in which accurate judgments are likely to occur.

Necessary Conditions (Realistic Accuracy Model)

Funder formulated a Realistic Accuracy Model of personality judgment (1999, also see Karing, Dörfler, & Artelt, 2013), which includes the necessary steps for judgments to be accurate. Within the model, four conditions are described which need to be fulfilled in order for accurate judgments to occur: 1) relevance; 2) availability; 3) detection; and 4) utilization.

The first necessary condition for accurate teacher judgments is *relevance*. In order to be able to arrive at an accurate judgment of a student's characteristics, the student has to reveal some kind of information that is potentially informative about the respective characteristic (e.g., answering a question posed by a teacher in math class or solving tasks on a science exam).

The second condition is *availability*, meaning that relevant information must be available to the teacher. For example, a student's inability to comprehend an aspect of the test instructions will have a direct impact on his or her test result, yet it may not be clear to the teacher that the student's poor performance is due to a misinterpretation rather than a cognitive failing. Therefore, unless the teacher takes further diagnostic steps, the information will remain unavailable to him.

The third condition of the model is *detection*, with teachers having to detect the relevant and available information. The detection may not always be conscious, but the informative stimulus must register in some way with the observer's nervous system. The detection of relevant information can be disturbed by different factors, such as distraction or inattentiveness (e.g., the teacher fails to register a student's cheating because of disruptions caused by other students).

Finally, the last condition and prerequisite for accurate judgment is *utilization*. Teachers have to correctly utilize the relevant, available, and detected information, interpreting it accurately, in terms of what it implies about the child's competence/achievement. Teachers can also (falsely) interpret a specific performance as a competence deficit, although poor performance in a given situation may be primarily due to a loss of interest and effort. In this respect, judging teacher's epistemological beliefs and stereotypes are relevant, since they might serve as schemes for specific attribution patterns. Overall, the model's four conditions are connected multiplicatively. Funder (1999) assumes that failure at any of the steps or in any of the conditions will lead to inaccurate judgments.

Variations of Judgment Accuracy Against the Background of the Realistic Accuracy Model

Funder's model (1999) describes mechanisms that help to explain some of the reported results on moderators of judgment accuracy. As described above, in their recent meta-analysis, Südkamp and colleagues (2012) found that teachers' judgment accuracy was moderated by use of either informed or uninformed teacher judgments, with use of informed judgments leading to a higher correspondence between teachers' judgments and students' academic achievement. Judgment accuracy thus differs depending on whether or not teachers are, for example, informed about the standard of comparison or the test used for the comparisons of their judgment with student characteristics. If they are aware of the contents of the test/tasks the students have to perform, the judgments tend to be more accurate. If the standards or the basis for the comparison is unknown to the teachers, however, they cannot be expected to detect relevant information that is potentially informative about the respective characteristic (relevance) nor are they able to decide whether relevant information is available to them (availability). The missing standard or knowledge about the problem's exact requirements also applies to both the detection and utilization steps in Funder's model (1999). Accordingly, the "inaccuracy" of teachers' judgments may be grounded in the studies' methodologies, rather than in the teachers' diagnostic competence, as Südkamp and colleagues (2012) themselves interpret the results.

Yet another finding is the lower accuracy of teacher judgments on students' affective and conative characteristics (e.g., Spinath, 2005; also found in our own research: Karing, 2009), which can be seen in light of this model (Karing et al., 2013). Relevant aspects of the underlying construct are not directly available to the teachers (e.g., task irrelevant thoughts in the case of test anxiety) and teachers have fewer opportunities – and are often not trained – to be attentive in this respect.

Necessary Information and Global Versus Task-Specific Judgments

What are the implications of the model of necessary conditions for the interpretation of our findings of judgment accuracy on either a global or a task-specific level

(Table 1)? Obviously, more information is available to the teachers for task-specific judgments. For global judgments, however, teachers have to infer which of the students' specific behavior they consider relevant for their judgment (detection). They also have to decide which detections of specific behavior they integrate in their overall judgment – which is also a potential source of error. For task-specific judgments, relevant information is available in the form of information about task demands. The detection and utilization of the available information, however, is dependent on the teachers' additional knowledge about the relative difficulty of the tasks (didactic knowledge), as well as their knowledge about students' individual strengths and difficulties relative to these task demands.

While the standards of comparison are quite clear for task-specific judgments, this is not necessarily the case for global judgments. Given that teachers have to infer their ratings based on their definition of reading proficiency or mathematical competence, the lower accuracy of global judgments for reading comprehension are possibly due to the lack of a common definition in this domain. Whereas reading proficiency is sometimes regarded mainly as a decoding skill and a kind of fluency indicator, other definitions (i.e., those inherent in the test used as the standard of reference in our study) are based on the idea of constructing a coherent mental representation of the text. Depending on what teachers consider to be the major part of the definition, their ratings may differ not because of a lack of judgment accuracy, but because of a mismatch between the test's inherent construct and the construct in the mind of the teachers. For the domain of mathematics, demands and competence models are clearer. Given that mathematical skill is mainly acquired in school and that the curriculum is well-defined according to age, teachers quite likely have a shared understanding of what constitutes mathematical proficiency. At the same time, this is likely to be the concept implemented in the reference tests for students, leading to higher judgment accuracy for mathematics as compared to reading. However, when looking at the difference between task-specific and global judgments, it should be remembered that the benefit of these task-specific judgments (as reported in Table 1), in terms of concrete and available information co-occurs with high cognitive demands on behalf of the teachers. Not only do they have to know the specific demands of the individual tasks students have to solve, but they also have to utilize their knowledge about the students' specific strengths and weaknesses in the context of the task. Clearly, then, content as well as didactical knowledge is a crucial element in this process.

Expertise – Knowledge

Teacher judgments about student characteristics relate, more or less, to specific school subjects. The more specific the ratings are in terms of a school subject or competence domain, the more likely it is that knowledge about the nature and demands of the domain is needed on the teacher's behalf in order for them to be able to judge students' performance adequately. One finding from our own research (Lorenz

& Artelt, 2009) can be interpreted in this respect: The accuracy of primary school teachers' ratings correlated substantially within a particular domain (e.g., between reading comprehension and vocabulary), whereas the correlation between domains (e.g., between mathematics and reading comprehension) was not significant.

Prior knowledge is a central feature of expertise; it is not, however, its only feature (Ericsson, Charness, Hoffman, & Feltovich, 2006). While there are a number of ways of operationalizing expertise, a quite common way is to use as an indicator the number of years a teacher has been practicing. As has been shown repeatedly, this variable (number of years teaching, above or below 5 years) does not seem to moderate the results of teachers' judgment accuracy. With respect to indicators of teachers' prior knowledge, however, the research evidence is very limited. While it is considered to be highly likely, we do not know yet, whether a detailed knowledge of the subject domain (content knowledge), in didactics (pedagogical content knowledge), and/or diagnostics, is predicative of teacher judgment accuracy. Given that knowledge of this kind is highly relevant for teacher training, it is quite surprising that the empirical evidence is still lacking. We have addressed this question in our own project of diagnostic competence, in the domain of text comprehension. Based on psychological, linguistic, and didactic theories regarding text and task difficulty, as well as the cognitive processes of solving text comprehension tasks (see Matthäi, 2012), a test was developed, that aimed at quantifying content knowledge and pedagogical content knowledge in this domain. Because data coding and analysis are still being carried out, we are not yet able to report the degree to which teachers' judgment accuracy is affected by the level of knowledge in the respective domain.

There are good reasons to assume that teachers' knowledge about tendencies and errors in the decision making process is a central element of a professional knowledge base and a theoretically relevant predictor of judgment accuracy. There are several tendencies and errors in teacher judgments that are well documented in the literature. Among these are errors related to selective information processing (expectancy effects, pygmalion effect), errors related to a false relying on dominant features (halo-effects), as well as errors related to naive personality theories, judgment tendencies and asymmetries in the attribution process (e.g., fundamental attribution error). In addition, errors occurring in everyday teacher assessment can be partly attributed to position (primacy/recency effects) and contrast effects, or they can be regarded as reference errors.

Judgment Relevance and Expertise

Krolak-Schwerdt and colleagues (e.g., Krolak-Schwerdt, Böhmer, & Gräsel, 2009, 2012) presented an important idea related to the role of teacher expertise for judgment accuracy. Based on social cognition theory (e.g., Fiske & Taylor, 2008), they argued that category-based and attribute-based decision making should be differentiated. Whereas attribute-based judgments are derived from the analysis of individual traits, category-based judgments rely on the processing of stereotype information. The resource-intensive attribute-based judgments are formed when the processing goal is accuracy (e.g., important decisions). This stands in contrast to the category-based judgments, which occur when the goal is to form a quick and efficient decision (e.g., impression formation). The more demanding processing modus is therefore only performed when the area is relevant and a rational decision is needed. Krolak-Schwerdt and colleagues have shown in multiple studies that teachers of varying experience (e.g., operationalized by years of teaching experience) differed in their ability to produce accurate judgments (including being able to switch the processing mode), if the goal of the decision making process was not only impression formation, but an accurate and professional judgment. For highly relevant decisions, expert teachers seemed to use resource intensive decision making (integrating contradictory issues, etc.), whereas novice teachers did not differ in their approach for decisions of varying importance. One of the limitations of the approach is the fact that neither the process of decision making was studied directly nor can ecological validity be assumed because the designs used were experimental and the teachers did not judge their own students but virtual students in a laboratory setting. Despite this limitation, the underlying idea of varying accuracy that depend on the subjective relevance of the result of the decision making process is well founded.

A similar – although more general – idea was developed by Kahneman (2011), who differentiates two ways the brain forms thoughts, called System 1 and System 2. According to Kahneman, System 1 produces fast, automatic, emotional, stereotypical, and subconscious thoughts and decisions, whereas System 2 consists of slow, effortful, infrequent, logical, calculating, and conscious thoughts and decisions. Kahneman (2011) assumes that System 1 is the predominant state of mind, whereas System 2 is only active on demand, that is, when the person decides to switch to active processing. Given that System 2 thinking is a very resource-intensive cognitive endeavor, it is quite likely that it is applied only when the person is motivated. The effect of motivation (and knowledge) on the processing mode is also highlighted by Hatano (1998). He suggests that the use of comprehension-oriented forms of learning is very time- and effort-intensive (high-cost, but high-benefit). In contrast to schema-based and automated cognitive information processing, comprehension-oriented learning involves the explicit checking of relations between previously acquired knowledge and new information, the formulation of hypotheses about possible connections, and the testing of these against the background of the new material. A learner is only assumed to engage in comprehension-oriented learning when the advantages of deploying such methods are anticipated to outweigh the disadvantages. Therefore, deeper levels of comprehension are closely related to higher levels of motivation. According to Hatano (1998), the assumed selectivity applies to the use of deeper processes of comprehension, as well as to the identification of gaps in one's own comprehension. There is no such tendency to overlook gaps in comprehension when one's own domains of interest or areas of expertise are involved. In that instance, detailed prior knowledge is available and serves as a basis for comprehension.

In the context of Kahneman's and Hatano's considerations, it is likely that accurate decision making is often limited to subjectively relevant and significant diagnostic decisions (or those that are extrinsically motivated). In addition, a profound knowledge base is instrumental in this process.

VULNERABILITY TO BIAS AND PREDICTION OF STUDENT PERFORMANCE FOR TASK-SPECIFIC AND GLOBAL JUDGMENTS

As mentioned previously, in the context of the BiKS Study, we asked teachers to judge student performance, either on a global level or by using task-specific judgments. While we do not know the degree of relevance teachers attach to either of these judgment types during the judgment task, the indicators seem to differ with respect to a few characteristics, which may lead to differential validity. As described above, the level of prior content knowledge with respect to the domain, as well as didactic and diagnostic knowledge, is particularly relevant for task-specific judgments. Therefore, we believe that the difference between these two judgment types lies not only in the availability of information about the test or the standard of performance, but also in the underlying judgment process: more precisely, in the cognitive demands inherent in this process. Given that estimating student performance on a specific task presupposes the integration of knowledge regarding task difficulty and student ability, task-specific judgments appear to be more demanding than global judgments; this assumption is in line with the pattern of results depicted in Table 1.

Vulnerability to Bias: Global Versus Task-Specific Judgments

Another theoretical difference between the two judgment types was discussed in a recent paper by our research group (Rausch, Karing, Dörfler, & Artelt, submitted). Dipboye and Gaugler (1993) have shown that unstructured judgment processes are more vulnerable to bias than those that are structured. In a similar vein, we hypothesized that, according to the inherent cognitive demands and the necessity to integrate different knowledge components, the influence of task-irrelevant aspects on judgments is less likely for task-specific (structured) than for global (unstructured) judgments. Therefore, global judgments were assumed to be more vulnerable to bias. Rausch and colleagues operationalized bias by estimating the additional impact of task-irrelevant aspects with students' achievement level controlled and found different results for both types of judgment. Bias can be found for global judgments (in the domains of mathematics and reading), but not for task-specific judgments. Teachers tend to overestimate the competence of students that are similar to themselves in terms of their personality profile (Big Five-based personality measure for both students and teachers). For both domains, personality similarity has a significant impact on teachers' global judgments, even when students' achievement levels are controlled for. Students who are more similar to their teacher are judged more positively than students who are more dissimilar. In neither domain, however, task-specific judgment is significantly influenced by personality similarity between student and teacher. Rausch and colleagues interpreted the effect of personality similarity on teacher judgment accuracy as a form of the similar-to-me effect (Rand & Wexley, 1975). According to this effect, similarity leads to sympathy, which might account for the positive judgments on students' competence level (overestimation). As indicated, this bias can only be found for global judgments. It therefore seems that task-specific judgments are less sensitive to bias since, due to the nature of the task, teachers are more likely to integrate knowledge of the content domain (e.g., task difficulty), as well as knowledge about the competence level of particular students. Teachers are forced to think about the way students accomplish particular tasks, presumably applying attribute based as opposed to category based processing.

Student Learning Progress and Teacher Judgment Accuracy

Teachers' diagnostic competence is assumed to be an important predictor of adaptive/tailored instructions, and therefore indirectly relevant for student learning progress (e.g., Brunner et al., 2011). In order to meet different students' needs, to provide accurate feedback and to tailor instructional designs accordingly, teachers need to be able to correctly judge relevant student characteristics, as well as the difficulty level and demands of class material (see also Anders et al., 2010). Since task-specific ratings cover both, judgments on student performance and task-difficulty, they are assumed to be a better predictor for students' learning gains than global judgments.

Altogether, only a few studies address the relevance of teachers' judgment accuracy on students' learning gains but, again, the pattern of results is heterogeneous. Schrader (1989) used rank-order indicators based on teacher judgments of either student ability or task difficulty, and found no direct effects of teachers' judgment accuracy for students' learning gains in mathematics. However, effects were found for teachers that used structuring aids during lessons, as well as adaptive teaching methods. Anders and colleagues (2010; see also Brunner et al., 2011) found direct effects on students' learning gains in mathematics using task-specific teacher judgments (rank-order indicators, as well as the task-related judgment error [mean absolute difference between empirical and estimated amount of correct responses per class]). For indicators of judgment accuracy based on global judgments, there seems to be no study that has reported direct effects on students' academic / educational outcomes. Aiming at estimating whether inter-individual differences in teacher judgment accuracy (measured with the most specific and knowledge-intensive indicator – the task-specific hit rate) affect students' competence development, Karing, Pfost, and Artelt (2011) analysed its effects on student performance, using data from the BiKS Study. Additionally, we tested whether this relationship was moderated by instructional variables. Teacher judgment accuracy was measured by a task-specific hit rate and the rank-order component in the domains of reading and mathematics. Multilevel analyses revealed that the task-specific hit rate was positively related to the development of students' reading competence, but it was not related to the development of students' mathematical competence. Furthermore, the significant relationship of teachers' task-specific hit rate was moderated by instructional variables such as their use of individualization and structuring cues. For the rank-order component, no significant positive relationships or interactions were found in the domains of reading and mathematics.

SUMMARY AND IMPLICATIONS

Teachers' diagnostic competence is a multifaceted construct. Until now, no common dimension has been identified that underlies the quality of decision making in different competence domains or different assessment purposes. The assumptions about an underlying competence (trait) are less pronounced, whereas at the same time, a state-like (vs. trait-like) interpretation widens the perspective so that situational (judgment specific) influences on teachers' judgments are more easily to be taken into account. Furthermore, against the background of a broader conceptualization of diagnostic competence (e.g., van Ophuysen, 2010; see also Schrader, 2011), judgment accuracy seems to be an adequate term, given that research is often limited to this facet of diagnostic competence.

Teachers' professional judgments clearly do vary with respect to the underlying purpose. Given the high number of formal and even higher number of informal decisions teachers have to make in everyday school context, it is unlikely that every decision is as accurate as it could be. Teachers need to make many decisions throughout the working day "on the fly", and thus, do not have the time to engage in an explicit decision-making process. That is, not every decision needs to be as precise as possible; given that the goals for some of the decisions are approximations or first impressions, for which category based judgments (according to social cognition theory), System 1 processing (according to Kahneman (2011)) or even fast and frugal decision making (according to Gigerenzer and Todd (1999)) seem to be well suited. However, as argued throughout this chapter, professional pedagogical decisions cannot rely on these heuristics. As was shown, vulnerability to bias is higher for unstructured, global decisions, whereas task-specific ratings of students' performance are not only less sensitive to bias, but also more predictive of students' future learning gains. If decisions are considered to be important by the teacher, such as when they are fundamental for the students' individual careers (like selection decisions for specific school types or tracks) or when important training decisions are based on these judgments, teachers need to be able to form as rational a decision as possible. This includes awareness of the possible threats and judgment errors, as well as an adequate way of dealing with them.

So far, research on judgment accuracy has produced little extractable knowledge for teacher training and professionalism. In this respect, it would be promising for further research to concentrate on construct-relevant rather than construct-irrelevant aspects of judgment accuracy. It seems of limited value to know that teachers who have a restricted understanding of the implemented concepts regarding student assessment are less accurate in their judgments of student characteristics than teachers who are informed about the assessment tasks. More research is needed on construct-relevant aspects. For example, to answer the question of which knowledge domains teachers need to learn in order to be able to form correct judgments about both student abilities and task demands. In this respect, a task-specific indicator that includes teachers' assessments of student competencies in relation to specific task demands is promising. Teachers are forced to think about the ways in which students accomplish particular tasks, presumably applying attribute-based as opposed to a category-based processing. Furthermore, task-specific judgments allow for detailed analyses of differential judgment accuracy, for example, varying by task difficulty, task demand, or task type. These kinds of findings are valuable for teacher training and other attempts at fostering teacher judgment accuracy. In order to deliver accurate judgments, teachers not only have to know the specific demands of the individual tasks students have to solve, but they also have to integrate their knowledge about students' specific strengths and weaknesses relative to these demands. Therefore, both content and didactical knowledge are crucial elements in this process. It makes sense to broaden the perspective on judgment accuracy by taking into account the process of teachers' diagnostic decision making: what kind of information (about students or tasks) is regarded as relevant for a judgment, and when does the process of (implicit) decision making become attribute based (rational diagnostic) processing? It seems that accurate decision making is often limited to subjectively relevant and significant diagnostic decisions. It often requires conscious processing of attributes (of tasks and students) and is more likely to occur when professional knowledge is available and when people are motivated to judge accurately.

NOTES

[1] The notion of diagnostic competence is widely used within the German-speaking countries. However, the implications and assumptions related to this term also apply to the notion of assessment literacy, a term that is more common in the Anglo-Saxon literature.

[2] These indicators of over- vs. underestimation as non-hit rate are similar but also are more specific than the indicators called judgment error or judgment tendency used by Anders et al. (2010) and McElvany et al. (2009) since they are estimated by simply calculating the difference between the estimated and empirically determined value of the corresponding student characteristic.

[3] BiKS is the German acronym for "Educational processes, competence development and selection decisions in preschool- and school age". It is an interdisciplinary research group funded by the German Research Foundation (DFG), and consists of eight research projects, one of which focuses on teachers' diagnostic competence (Artelt & Weinert - AR 301/6-1, AR 301/ 6-2, AR 301/ 6-3).

REFERENCES/BIBLIOGRAPHY

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 its impact on students' achievement]. *Psychologie in Erziehung und Unterricht, 57*, 175–193.

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

Brunner, M., Anders, Y., Hachfeld, A., & Krauss, S. (2011). Diagnostische Fähigkeiten von Mathematiklehrkräften [Mathematics teachers' diagnostic skills]. In M. Kunter, J. Baumert, W. Blum, U. Klusmann, S. Krauss, & M. Neubrand (Eds.), *Professionelle Kompetenz von Lehrkräften. Ergebnisse des Forschungsprogramms COACTIV* (pp. 215–234). Münster: Waxmann.

Coladarci, T. (1986). Accuracy of teacher judgments of student responses to standardized test items. *Journal of Educational Psychology, 78*, 141–146.

Demaray, M. K., & Elliott, S. N. (1998). Teachers' judgments of students' academic functioning: A comparison of actual and predicted performances. *School Psychology Quarterly, 13*, 8–24.

Dipboye, R., & Gaugler, B. B. (1993). Cognitive and behavioral processes in the selection interview. In N. Schmitt & W. C. Borman (Eds.), *Personnel selection in organizations* (pp. 135–170). San Francisco, CA: Jossey-Bass.

Ericsson, K. A., Charness, N., Hoffman, R. R., & Feltovich, P. J. (Eds.). (2006). *The Cambridge handbook of expertise and expert performance.* Cambridge, MA: Cambridge University Press.

Feinberg, A. B., & Shapiro, E. S. (2009). Teacher accuracy: An examination of teacher-based judgments of students' reading with differing achievement levels. *Journal of Educational Research, 102*, 453–462.

Fiske, S. T., & Taylor, S. E. (2008). *Social cognition: From brains to culture*. New York, NY: McGraw-Hill.

Funder, D. C. (1999). *Personality judgment. A realistic approach to person perception*. San Diego, CA: Academic Press.

Gigerenzer, G., & Todd, P. M. (1999). *Simple heuristics that make us smart.* New York, NY: Oxford University Press.

Hatano, G. (1998). Comprehension activity in individuals and groups. In M. Sabourin, F. Craik, & M. Robert (Eds.), *Advances in Psychological Science. Volume 2: Biological and cognitive aspects* (pp. 399–418). Hove, UK: Psychology Press/Erlbaum.

Helmke, A., Hosenfeld, I., & Schrader, F. W. (2004). Vergleichsarbeiten als Instrument zur Verbesserung der Diagnosekompetenz von Lehrkräften [Comparisons of instruments to foster teachers' diagnostic competency]. In R. Arnold & C. Griese (Eds.), *Schulleitung und Schulentwicklung* (pp. 119–144). Hohengehren: Schneider Verlag.

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

Kahneman, D. (2011). *Thinking, fast and slow*. New York, NY: Farrar, Straus and Giroux.

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.

Karing, C., Dörfler, T., & Artelt, C. (2013). How accurate are teacher and parent judgements of lower secondary school children's test anxiety? *Educational Psychology*.

Karing, C., Matthäi, J., & Artelt, C. (2011). Genauigkeit von Lehrerurteilen über die Lesekompetenz ihrer Schülerinnen und Schüler in der Sekundarstufe I – Eine Frage der Spezifität [Lower secondary school teacher judgment accuracy of students' reading competence – A matter of specificity]? *Zeitschrift für Pädagogische Psychologie, 25*, 159–172.

Karing, C., Pfost, M., & Artelt, C. (2011) Hängt die diagnostische Kompetenz von Sekundarstufenlehrkräften mit der Entwicklung der Lesekompetenz und der mathematischen Kompetenz ihrer Schülerinnen und Schüler zusammen? [Is there a relationship between lower secondary school teacher judgment accuracy and the development of students' reading and mathematical competence?] *Journal for Educational Research Online. (JERO), 3*, 119–147.

Karst, K. (2012). *Kompetenzmodellierung des diagnostischen Urteils von Grundschullehrern [Competence modeling of primary school teachers' diagnostic judgment]*. Münster: Waxmann.

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

Krolak-Schwerdt, S., Böhmer, M., & Gräsel, C. (2012). Leistungsbeurteilungen von Schulkindern. Welche Rolle spielen Ziele und Expertise der Lehrkraft? [Students' achievement judgments: The role of teachers' goals and expertise] *Zeitschrift für Entwicklungspsychologie und Pädagogische Psychologie, 44*, 111–122.

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.

Matthäi, J. (2012). *Was beeinflusst das Textverstehen? Theoretische Erkenntnisse und Perspektiven für die Praxis*. Bamberg: Universität Bamberg.

McElvany, N. et al. (2009). Diagnostische Fähigkeiten von Lehrkräften bei der Einschätzung von Schülerleistungen und Aufgabenschwierigkeiten bei Lernmedien mit instruktionalen Bildern [Teachers' diagnostic skills to judge student performance and task difficulty when learning materials include instructional pictures]. *Zeitschrift für Pädagogische Psychologie, 23*, 223–235.

Rand, T. M., & Wexley, K. N. (1975). Demonstration of the effect, "similar to me," in simulated employment interviews. *Psychological Reports, 36*, 535–544.

Rausch, T., Karing, C., Dörfler, T., & Artelt, C. (submitted). Personality similarity between teachers and their students influences teacher judgment of student achievement. (Manuscript submitted for publication).

Schrader, F. W. (2010). Diagnostische Kompetenz von Eltern und Lehrern [Diagnostic competence of teachers and parents]. In D. H. Rost (Ed.), *Handwörterbuch Pädagogische Psychologie* (pp. 102–108). Weinheim: Beltz Verlag.

Schrader, F. W. (2011). Lehrer als Diagnostiker [Teachers as diagnosticians]. In E. Terhart, H. Bennewitz, & M. Rothland (Eds.), *Handbuch der Forschung zum Lehrerberuf* (pp. 683–698). Münster: Waxmann.

Schrader, F.-W. (1989). *Diagnostische Kompetenzen von Lehrern und ihre Bedeutung für die Gestaltung und Effektivität des Unterrichts [Teachers' diagnostic competence and its significance for the presentation and effeciency of instruction]*. Frankfurt: Lang.

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

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.

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.

van Ophuysen, S. (2010). Professionelle pädagogisch-diagnostische Kompetenz – eine theoretische und empirische Annäherung [Professional pedagogical-diagnostic competence – A theoretical and empirical approach]. In N. Berkemeyer, W. Bos, H. G. Holtappels, N. McElvany, & R. Schulz-Zander (Eds.), *Jahrbuch der Schulentwicklung. Daten, Beispiele und Perspektiven* (Vol. 16, pp. 203–234). Weinheim: Juventa.

AFFILIATIONS

Cordula Artelt
Department of Educational Research,
University of Bamberg

Tobias Rausch
Department of Educational Research,
University of Bamberg

INEKE PIT-TEN CATE, SABINE KROLAK-SCHWERDT,
SABINE GLOCK & MARIA MARKOVA

IMPROVING TEACHERS' JUDGMENTS: OBTAINING CHANGE THROUGH COGNITIVE PROCESSES

A central aspect of teachers' professional competence is the ability to judge students' achievements adequately. Giving grades and marks is the prototypical task in this context. Besides giving grades, assessments for school placements or tracking decisions belong to these tasks. Other judgments are more implicit in that no specific judgment is required, but students' achievements are estimated intuitively. Examples are decisions made during class such as "calling on a particular student". These judgments have substantial relevance for individual students, and consequently, high competence in judging students correctly is seen as a key skill for teachers and future teachers (Shepard, 2006). However, at the same time, a number of studies have shown that teachers' judgments of student performance frequently do not meet the criteria of measurement theory such as reliability and validity, but seem to be rather subjective (Givvin, Stipek, Salmon & MacGyvers, 2001; Swanson, 1986). Within educational systems where judgments are used to make decisions about a student's future academic career, this may contribute to problems of social segregation and may be harmful to the personal and later professional development of students (Alpert & Bechar, 2008).

In their meta-analysis on teacher judgment accuracy Hoge and Coladarci (1989), and more recently Südkamp, Kaiser and Möller (2012), come to the conclusion that although teachers' judgment accuracy of students' performance is fairly high, the teachers' judgments leave 57% up to 72% of the variation of students' test performance unexplained 'which leaves plenty of room for improvements' (Südkamp et al. 2012, p. 13). In this regard, research has consistently provided evidence that, although academic achievement is important, teachers' judgments and decision making processes are also influenced by non-academic variables such as the social and migrant background of students. To analyse teachers' judgments, insights from the field of social judgment formation and decision making have proven valuable. This theoretical framework focuses on the question how person attributes, such as behaviours, beliefs, etc., are selected and integrated into a judgment. Applied to education, the questions concern how teachers select, use and integrate student information, such as grades, gender, background, motivation and behaviour, into a judgment. Theories of social judgment formation consider a decision as the result of a cognitive process involving not only the search for information, whereby one has to decide on which type information is to be acquired, but also the application

S. Krolak-Schwerdt et al., (Eds.), Teachers' Professional Development, 45–61.

of (implicit) rules regarding the use of information. Teacher expectations have been identified to affect this cognitive processes, such that teachers' stereotypical expectations about students' achievements on the basis of socioeconomic or ethnic background, or gender affects teacher judgment (e.g. Andrews et al. 1997; Brophy & Good, 1974; Parks & Kennedy, 2007; Pigott & Cowen, 2000; Reyna, 2000; Weiner, 2000). In addition, several variables have been identified as moderator variables such as teachers' goals, motivations, and accountability. Biases in judgments due to expectations are more likely to occur when there is an incentive to confirm an expectation or a striving to rapidly reach a particular conclusion. Judgment biases are less likely when there is motivation to develop an accurate impression of the target person or when the perceiver's outcomes depend on the target person (see Jussim, Eccles, & Madon, 1996, for a review). For example, teachers' assessments of students' achievements become less biased when teachers have the goal of improving students' achievements (Goldenberg, 1992) or when assertive parents offer evidence that conflicts with teachers' expectations (Good & Nichols, 2001).

Assuming an association between teachers' cognitions and student learning (see Orton, 1996), changing teachers' cognitions may then improve student performance. In order to change teachers' cognitive processes, teachers have to be informed about the processes, which might unconsciously influence their classroom behaviour and judgments. Knowledge concerning the different processes and their consequences enables teachers to counter these effects. To this extent, there are several phenomena of which teachers should be aware to avoid unconscious influences. This chapter will outline the extent to which teachers' cognitions and beliefs may affect teachers' judgments, and their association with student learning. Moreover, we will focus on factors that can moderate teachers' cognitive processes and on trainings to improve the quality of teachers' judgments. More specifically, following the above we will first focus on teacher expectations, and then show how accountability could moderate teachers' cognitive processes. Finally we will discuss how statistical prediction rules, which confront decision makers with immediate feedback on the relation between predicted and actual decisions, may be utilised to reduce bias and errors in decision making and hence improve teacher judgment accuracy.

TEACHER EXPECTATIONS AND STEREOTYPE THREAT

Since the pivotal study from Rosenthal and Jacobson (1968) teacher expectations are discussed as important factors that influence teachers' judgments and students' academic performance. More specifically, teachers interact with their students corresponding to their expectations and this behaviour might lead their students to act consistent with these expectations (Brophy & Good, 1974; Rosenthal & Jacobson, 1968). Although expectancy effects are rather small and nowadays teachers should be informed about them (Jussim & Harber, 2005), they persist to occur in the classrooms. This suggests that merely informing teachers is not enough to ensure that such effects could be avoided. Rather, to avoid expectancy effects in the classroom, teachers have to be made aware of their expectations regarding their students and understand where these stem from. Although generally teachers are not told what they have to expect of a particular student, they nevertheless hold expectations, which often stem from stereotypes, which may be conceptualized as knowledge about members of social groups (e.g. Fiske & Taylor, 1991). These knowledge structures simplify the world, in that people use them in judgment formation (e.g. Macrae & Bodenhausen, 2000; Stangor & Schaller, 1996). With experience, teachers learn much about different students and develop stereotypes about students who share important characteristics. Based on these stereotypes, expectations develop which can colour perception and judgments (Ferguson, 2003). However, in order for teacher expectancy effects to occur as a result of self-fulfilling prophecies, students' actual academic performance needs to adapt to the teachers' expectations (Jussim & Harber, 2005). In other words, teachers not only have to hold expectations and act accordingly, but students also have to react consistent with these expectations. In this process, stereotype threat (Steele, 1998; Steele & Aronson, 1995) comes into play. Stereotype threat is the phenomenon that academic achievement decreases among the members of negatively stereotyped groups due to the fact that the intellectual capacity of the group is assumed to be low (Steele, 1998). This decrease in academic achievement is due to anxiety of the group members because they know about the negative stereotypes and do not only risk personal failures, but also risk confirmation of the negative intellectual group stereotype (Osborne, 2001; Steele, 1998). There is ample evidence that members of intellectually stigmatized groups actually perform lower in achievement tests, particularly when their group membership is salient (e.g., Aronson et al., 1999; Steele & Aronson, 1995; Steele, Spencer, & Aronson, 2002). Although in theory positive associations with stereotype expectations are feasible in the way that students might profit from high teacher expectations, most research has focused on the fact that students from minority groups are more susceptible for low teacher expectations (McKown & Weinstein, 2002) and react with stereotype threat.

ATTITUDES, STEREOTYPES, AND SUPPRESSION

Teachers should prevent that stereotypes affect their own judgments. Südkamp et al. (2012) suggested teachers' judgments may depend on stereotypes, which could affect accuracy. To this extent, research has provided evidence for stereotypes biasing teachers' judgments (Krolak-Schwerdt, Böhmer, & Gräsel, 2012), particularly racial stereotypes coloured teachers' judgments (McCombs & Gay, 1988; Neal, McCay, Webb-Johnson, & Bridget, 2003; Parks & Kennedy, 2007). Even student teachers' judgments were already biased through ethnicity (Glock & Krolak-Schwerdt, 2013). Particularly, when members of social groups strongly confirm stereotypical expectations, person perception and judgments are coloured through stereotypes (Stangor & McMillan, 1992). Thus, minority students who behave like typical exemplars of the stereotype are at risk to get stereotyped (e.g., Glock & Krolak-Schwerdt, 2013; Neal et al., 2003). Moreover, not only stereotypes shape person perception and judgment, but also attitudes (Sanbonmatsu & Fazio, 1990). Usually, a distinction between explicit and implicit attitudes is made (Gawronski & Bodenhausen, 2006). Explicit attitudes are thoughtful reflections (Gawronski & Bodenhausen, 2006) people engage in to derive an evaluation of an attitude object. The expression of explicit attitudes involves controlled and effortful processes (Fazio, 1990; Gawronski & Bodenhausen, 2006), because people have to either construct an evaluation on the spot (Bassili & Brown, 2005; Schwarz & Bohner, 2001) or retrieve the evaluation from memory. Thus, the expression of explicit attitudes always depends on the ability and on the motivation to engage in those processes (Fazio, 1990; Fazio & Towles-Schwen, 1999). Relying on self-report measures, explicit attitudes are prone to social desirability bias (De Houwer, 2006). By contrast, implicit attitudes are automatic evaluations (Gawronski & Bodenhausen, 2006) that automatically come into mind whenever the attitude object is present (Fazio, 2007; Olson & Fazio, 2009). Especially implicit attitudes often guide behaviour, affect judgments, and determine how information is processed (Houston & Fazio, 1989; Schuette & Fazio, 1995). Thus, implicit attitudes are crucial, as they play a pivotal role in situations which are cognitively demanding and in which cognitive resources are restrained (Hofmann, Gschwendner, Castelli, & Schmitt, 2008). This might be of particular relevance for teachers in classrooms, as teaching can be stressful (van Dick & Wagner, 2001) and teachers are often required to manage excessive demands under time pressure (Santavirta, Solovieva, & Theorell, 2007).

Research on implicit attitudes towards minority students among teachers is sparse. Although there are some studies focusing on teachers' implicit attitude towards students with special needs (Enea-Drapeau, Carlier, & Huguet, 2012; Hornstra, Denessen, Bakker, van den Bergh, & Voeten, 2010; Levins, Bornholt, & Lennon, 2005) the paradigm has not been used often to explain teachers' judgments about students from ethnic minority groups. There is one study (van den Bergh, Denessen, Hornstra, Voeten, & Holland, 2010) which investigated implicit attitudes and their relation to teachers' judgments. The authors provided evidence for negative implicit attitudes towards minority students being a strong predictor of teachers' expectations and of their achievement judgments while explicit attitudes neither had a predictive value for achievement judgments nor a relationship to implicit attitudes (van den Bergh et al., 2010).

Although there is ample evidence for stereotypes affecting teachers' judgments and some evidence for the role of implicit attitudes in teachers' expectations and judgments, teachers could be trained in stereotype and implicit attitudes suppression. Training in stereotype and attitudes suppression could avoid the rebound effect, that is, the tendency of people to rely on stereotypes much stronger after suppression than before. Stereotype suppression is, when untrained, cognitively demanding and the resource depletion elicits stereotyping afterwards, because stereotyping occurs without much cognitive resources (Macrae & Bodenhausen, 2000). Thus, training teachers in stereotype suppression and in controlling their attitudes would result in automatic suppression, and this, in turn, would leave open cognitive resources, which could be used for instructional and classroom demands (Kawakami, Dovidio, Moll, Hermsen, & Russin, 2000). One way to ensure the success of training is to integrate feedback as a key factor. Empirical findings in the fields of instructional skills and diagnostic competence have demonstrated the importance of feedback in developing training programs to improve judgment accuracy. It should be noted that transfer of new skills to classroom behaviour has proven difficult, especially for experienced teachers, requiring the implementation of numerous feedback loops (Scheeler, 2008). In contrast, changing teachers' cognitive processes, which are associated with teaching behaviour and teachers' judgments, might be reached with relative ease (Wahl, Weinert, & Huber, 2007) and can also be transferred to school practice without difficulty (Helmke, Hosenfeld, & Schrader, 2004). Feedback at the cognitive level to improve judgment accuracy could utilize self-reflection, whereby teachers' predictions of student achievement are compared to the results of the students on standardized achievement tests (Wahl, et al., 2007). Via this method, teachers' implicit hypotheses and judgments are subject to explicit and empirical control, in which discrepancies between judgment and actual achievement could be consulted to find sources of errors.

In sum, many mechanisms might bias and influence the teachers' cognitive processes but there are possibilities to overcome these often automatic mechanisms, which result in more accurate and less biased teachers' judgments.

ACCOUNTABILITY

As stated above, cognitive processes and associated judgments might be influenced and biased by different factors. So far, we have focused on the extent to which stereotypical beliefs may bias teacher judgments. However, biases may also result from the way the teachers process the information upon which judgments are based.

In social cognitive psychology, theories of judgment formation have been put forward to describe and explain different ways in which people form judgments of other people. One group of models assumes that people collect information in a systematic way and weigh and integrate these informational cues when making a decision. Such information integrating strategies (e.g. Dawes & Corrigan, 1974; Brehmer, 1994; Swets, Dawes & Monahan, 2000) lead to deliberate decisions. Another group of models assumes less complex judgment processes: A judge relies on only a minimum of critical cues to make a decision (e.g. Gigerenzer & Todd, 1999; Hoffrage & Reimer, 2004) whereby stereotypes have priority and determine the nature of the judgment while other relevant cues are widely ignored (Bodenhausen, Macrae & Sherman, 1999; Fiske & Neuberg, 1990). Accordingly, such stereotype-based strategies pose judgment formation processes to be highly cognitive economical and efficient. Dual process theories of impression and judgment formation (Fiske & Neuberg, 1990) posit that people can shift between the two processing strategies in response to certain demands and in accordance with motivation. The stereotype-based strategy occurs when the available information about a target person easily fits already familiar stereotypes (Gilbert & Hixon, 1991). The information integration strategy mainly occurs when the actual information does not easily fit stereotypes or when people have high motivation and cognitive resources to engage in the processing of individual information. In this processing strategy, cues which are diagnostic for the judgment are collected in a systematic way, carefully elaborated and integrated into the decision. Research has shown that teachers shift between the two strategies, depending on the situational context, their goals, and their motivation. Krolak-Schwerdt and colleagues (Krolak-Schwerdt, Böhmer, & Gräsel, 2009; Krolak-Schwerdt et al., 2012) demonstrated that teachers involved in more thorough examination of students' profiles and were more likely to use the information integration strategy when they were asked to predict the student's future educational career. In contrast, teachers who were instructed to form an impression of the student subsequently relied more on available stereotypes.

In general, people who are highly motivated to be accurate, preferentially use the information integration strategy, whilst people with low motivation to attend to the given target person's information more likely rely on the stereotype-based strategy (Gollwitzer & Moskowitz, 1996; Kunda & Spencer, 2003; Quinn & Schlenker, 2002). Thus, the motivation moderates the activated information processing strategy. The motivation is influenced by the need to justify the judgment to a third party (Pendry & Macrae, 1996), which increases the accountability for the judgment (Tetlock, 1983). When people are made accountable towards an anonymous third party before they engage in the encoding of information and judgment formation, they make more use of the information integration strategy (Lerner & Tetlock, 2003). This is indicated by reduced levels of overconfidence (Siegel-Jacobs & Yates, 1996), the use of less traits in person descriptions (Boudreau, Baron, & Oliver, 1992) and reduced cognitive or judgment biases resulting from effort demanding, integrative complex and evaluative inconsistent thinking, required to demonstrate awareness of alternative perspectives (Tetlock, 1983). Increasing people's expectancies about personal consequences before they form a judgment initiates a need for accuracy, whereas receiving such information after a judgment is made generates fear of invalidity.

Being accountable for a judgment with serious consequences may also result in high attention to and careful integration of all available information (Lee, Herr, Kardes, & Kim, 1999; Lerner & Tetlock, 1999). Consequently, accountability – defined as people's implicit or explicit expectations to justify their beliefs, feelings, and actions to others (Tetlock & Lerner, 1999) – could be a moderator of judgment formation strategies. The findings of Krolak-Schwerdt and colleagues (2009, 2012) support the role of accountability as a moderating factor because having to predict the student's future academic career increases teachers' personal accountability for the judgment (Glock, Klapproth, Böhmer, & Krolak-Schwerdt, 2012; Glock, Krolak-Schwerdt, Klapproth, & Böhmer, 2012).

Generally, high accountability towards an external audience is associated with the consideration of more information, spending more time examining information, and the consideration of more alternative decisions than low accountability (Lee et al., 1999). High accountability further leads to increased depth and complexity of information processing, regardless of people's previous competence concerning the task (Lee et al., 1999). High accountability may draw people's attention towards extrinsic rewards, such as audience's approval of their decision (Lee et al., 1999).

However, there are contradicting empirical results concerning the effects of accountability. Some studies have shown accountability to amplify bias (Hattrup & Ford, 1995; Siegel-Jacobs & Yates, 1996), others have found no effect on the quality of people's judgments (Johnson & Kaplan, 1991; Simonson & Nye, 1992), whereas there are also studies reporting that increased accountability can attenuate bias (e.g. Tetlock, 1985; Thompson, 1995). One explanation of these mixed findings may be the use of different definitions of accountability or applied theoretical frameworks and research designs. Different experimental manipulations of accountability could also explain the mixed findings. According to Tetlock, Skitka, and Boettger (1989) people respond differently to accountability demands depending on the situation they find themselves in. For example, when people know the views of the audience, they shift their own views towards those of the prospective audience. In other words, they are likely to adopt the salient, socially acceptable position, as this saves cognitive work. In contrast, when people do not know the views of the constituency, they are motivated to think in relatively flexible, multidimensional ways and involve in self-critical information processing trying to anticipate the objections of potential critics. This accountability coping strategy is a process of pre-emptive self-criticism, which improves performance and stimulates information-integration processing but, at the same time, increases sensitivity to risk (Tetlock et al.,1989). Still another way to deal with demands of accountability is called defensive bolstering: People who expect to be held accountable for positions, to which they feel committed, devote the majority of their mental effort to justifying those positions (Tetlock et al., 1989). Finally, when people are accountable to conflicting audiences, when the potential risks of the judgment are moderate to high, and when it is necessary to impose losses in order to promote general good, people tend to engage in procrastinating and other judgment avoidance strategies (Tetlock et al.,1989).

Each one of those four coping responses has differential effects on judgment quality and could be adaptive in different circumstances. This could explain partially the contradicting empirical results on effects of accountability, that is, accountability leads to more use of information-integrating processing strategies and less bias only under certain circumstances. Most empirical evidence, however, puts forward bias-reducing effects of accountability on information processing. People who were held accountable for their judgments generally invested cognitive effort into making judgments and decisions (Tetlock, 1983), indulged in a deeper information search and spent more time to arrive at a decision (Hattrup & Ford, 1995). Lerner and Tetlock (1999, 2003) argue that the expectation of having to justify one's views motivates people to be more attentive information processors and increases the likelihood to perform difficult tasks, both indicators of high quality judgments. In addition, this processing strategy most likely increases judges' resistance towards different cognitive biases, particularly biases resulting from a reliance on stereotypes.

Most of the above referred accountability effects apply to educational context in which teachers are required to judge their students. Results of two recent studies (Glock, Klapproth, et al., 2012; Glock, Krolak-Schwerdt, et al., 2012) showed that teachers with low accountability were twice as likely to orient students without immigrant background to the highest secondary school track compared to ethnic minority students, even after controlling for achievement level. In contrast, teachers with high accountability did not differentiate between students with different ethnic backgrounds and similar achievement profiles, indicating that increased accountability is associated with reduced stereotypical bias. Studies have also demonstrated that a change of motivation led to less biased judgments. That is, in the case of school tracking recommendations, objectivity is improved when teachers receive pre-decisional accountability instructions (Glock, Krolak-Schwerdt, et al., 2012; Krolak-Schwerdt et al., 2009, 2012). In effect, just asking teachers about their perceived accountability already increased the accuracy of the judgments, thus leading to less biased judgments (Pit-ten Cate, Krolak-Schwerdt, Glock, & Markova, 2012). More specifically, after accountability priming, teachers' transition decisions became not only more accurate, but as a result of increased accountability, differences in accuracy of transition decisions for students from different backgrounds reduced.

These empirical findings confirm that accountability moderates the use of processing strategy. Applied to the educational context, research shows accountability differentially shifts teachers' processing of student information and their assessments of student performance, whereby teachers will shift from a stereotype-based to an information integrating strategy. More specifically, low accountability induces stereotype-based processing with stereotypes affecting attention, memory and judgment, whereas high accountability directs attention to the individual information given about a student with memory and judgment being unaffected by stereotypes. In addition, this line of research sheds light on the cognitive processes that underlie the variations in the quality of teachers' judgments by demonstrating that increased accountability influences early phases in the processing of student information, that is, attention and memory. This, in turn, may also constitute the cognitive mechanisms of relatively more biased or accurate judgment formation in the educational domain.

STATISTICAL PREDICTION RULES

As stated above, judgment accuracy may be affected by racial, social class, or gender bias. Bias may result not only from stereotypical beliefs but also from the way the judge integrates information upon which the decision is based (Garb, 1997). One way to improve judgments is to focus on diagnostic competence, that is, the skillset to judge people adequately. In education, diagnostic competence would entail judgments of students' academic achievement and would include the ability to formally assign grades for school work or to provide recommendations for school placement as well as the more intuitive and informal estimation of student performance and behaviour in class. Improving judgment accuracy may involve increasing the ability to distinguish between alternatives and to select the correct one. For example, if teachers have to judge students' academic performances, they will need to choose between the alternatives 'achieved' or 'not achieved', and possible intermediate levels. Similarly, if a teacher has to decide which type of education suits a student best, available information needs to be assessed and a choice is then made between different options (e.g. school tracks). However, such judgments may be prone to bias given underlying cognitive processes. An alternative way to increase judgment accuracy, especially in situations with fixed alternatives (yes-no; achieved-not achieved), may be to rely not only on improving accuracy, but also take into account the probability of the alternative decisions as well as the benefits and costs of the (in)correct decisions (Swets, et al., 2000). That is, judgment accuracy is affected by the extent to which different alternatives are possible given a certain student profile and therefore one should consider the consequences of the different outcomes for the student.

Given the risk of bias associated with judgments, especially when affected by intuitive inferences, judgment accuracy could be improved by using formal decision rules on the weighted integration of informational cues, which have a proven diagnostic value for the judgment. Such statistical prediction rules (SPR) can be created by aggregating relevant information about the issue to be judged (predictor variables) into a decision (Swets et al., 2000).

The use of SPRs in terms of linear models is not new. In a review, Dawes and Corrigan (1974) outlined the utility of linear models in decision making, dating the first normative use of linear models as far back as 1887. The universal use of linear models follows their appropriateness given the characteristics of various decision making situations. The authors concluded that linear models outperform intuitive judgments in situations in which the predictor variables have a conditionally monotone relationship to the criterion (e.g. no matter how students score on other variables, they are more likely to fare better when they score higher on a specific achievement test). Furthermore, linear models are not greatly affected by measurement error in the dependent variable, and possible measurement error in the predictor variables will tend to increase linearity (Dawes & Corrigan, 1974). Since then, numerous studies have shown linear models to be generally useful in modelling individual decisions in different areas, specifically in the field of medicine (e.g. Bankowitz, McNeil, Challinor, & Miller, 1989; Berner et al., 1994; Getty et al., 1997) and psychology (e.g. Grove, Zald, Lebow, Snitz, & Nelson, 2000). For example, Berner and colleagues (1994) found that computer generated medical diagnoses in 70-90% matched the clinicians list of possible diagnoses for 105 cases. In addition, newly generated computer diagnoses were retrospectively considered valuable by the clinicians. Getty and colleagues (1994) showed that the optimal integration of cues yielded an improvement in accuracy of prostate cancer staging, whilst Bankowitz and colleagues (1989) showed that SPRs could be a valuable tool to provide feedback to clinicians, as they showed clinicians felt inclined to change or considered changing their diagnosis after consulting the SPR predictions. In a meta-analysis, Grove and colleagues (2000) showed that in up to 47% of the studies they examined, SPR based judgments outperformed clinical predictions and that on average such mechanical judgments were 10% more accurate than clinical predictions. More recently, Aegisdóttir and colleagues (2006) concluded that clinical predictions of mental health practitioners were generally less accurate than predictions based on statistical methods (effect size -.12, i.e. accuracy levels increased by 13% when using statistical techniques rather than clinical judgment). The use of SPRs will address common problems associated with human decision making, such as bias. Also, SPRs may provide valuable feedback to clinicians which will give them insights into their decision making processes and allow them to change less effective habits (Garb, 1997; Grove et al., 2000).

Thus, findings suggest that SPR predictions may be more accurate than clinical judgments and some have recommended clinical judgments should therefore be replaced by SPRs. Others however, see SPRs more as a tool to guide clinicians in their decision making. In this respect, Brehmer and Brehmer (1988) reviewed research on the use of linear models and concluded that, although linear models generally fit judgments well, judges often use few cues and use them inconsistently. In addition, there are considerable inter-individual differences in the assigned weight to cues. Thus, linear models may prove effective in judgment situations in which it is standard practice to review different cues and rules on how to combine information, and when results are viewed in terms of accuracy, but are less useful when striving for uniformity, as the context in which judgments are made will result in inter-individual differences as to the selection and use of decision rules. Indeed, judgment accuracy depends on the correlation between the decision making rules and the environment (Hogarth & Karelaia, 2007). Linear models can, however, be fruitful as a feedback tool in a dynamic process of human decision making, as they will provide guidance to the judge (Brehmer, 1994). However, whilst accepting possible limitations of SPRs in certain situations, Dawes (2002) argued that if well validated SPRs, that generally outperform professional judgment are available, professionals should replace, rather than use to educate, one's intuitive judgment, especially within the psychological profession. To this extent, Dana and Thomas (2006) also commented that, although clinical expertise should not be dismissed, given the superiority of SPRs over human judgment, there are no grounds to refrain from using SPRs for socially important decisions.

In summary, the use of SPRs could lead to a higher consistency in judgments. This means that decision makers would make the same decision each time for any given set of information. This may be of particular importance when decisions are based on a combination of objective and subjective information and when the decision maker is more or less accountable for his/her decisions (e.g. within the field of education). SPRs can increase judgment accuracy, and may be most useful in supplying the judges with objective output which they can then use to make a final judgment. Within the educational domain, this approach may be especially useful when teachers make judgments for their students' schooling. These judgments may not only concern the short term (e.g. does the student need extra learning support?), but also the long term (e.g. can the student proceed to the next class or which secondary track would be most suitable for this student?). Given the importance of such judgments and the success of SPRs in other domains, one should encourage both teachers and student teachers to use SPRs in order to increase judgment accuracy by reducing bias and error.

CONCLUSION

The ability to make valid and reliable judgments of student achievement is a key component of teachers' professional competence. This chapter has focused on the role of cognitive processes in decision making. We have shown that changing cognitive processes associated with teachers' judgments affects teaching behaviour, resulting in a reduced influence of intuitive beliefs and stereotypes and increased accuracy. Methods have included stereotype suppression, goal-setting, and increased accountability, as well as the application of SPRs. Although different in nature, all approaches have in common that they aim to raise teachers' awareness of their intuitive inferences, to overcome stereotype bias in judgment and to reduce judgment discrepancies between individuals. Research has consistently shown that stereotype bias is, at least temporarily, reduced as a result of changing the underlying cognitive processes. More specifically, teachers are more inclined to consider a range of information rather than to rely on stereotypical beliefs in situations, in which they are motivated to suppress stereotypes, either by increased awareness, increased accountability, or the application of formal decision rules. So far most research has focused on the cognitive processes themselves (e.g. Glock, Klapproth et al. 2012; Krolak-Schwerdt et al 2012), that is, the association between differences in processing strategy and bias. However, limited data exist on the effect of this on judgment accuracy. To this extent, Jussim and Harber (2005) have commented that although social psychologists generally assume reduced bias will alleviate self-fulfilling prophecies in the classroom, educational research has shown that teacher expectancies are generally accurate. They concluded that more research is needed to investigate to what extent the validity of teachers' judgments creates, sustains or alleviated social injustices. Furthermore, limited information exists on the relative efficacy of the various methods used to accomplish change. Therefore, more research is needed to evaluate the effects of different modes of establishing changes in cognitive processes on judgment accuracy. In this evaluative process, the different techniques may be more or less suitable in different situations. For example, the extent to which accountability levels could be increased may be dependent on the school structure or educational system whereas the use of prediction rules may be especially useful for trainings, as they enable to provide teachers with cognitive feedback. Such situational circumstances should be taken into account when making recommendations for training.

First studies on the association between overcoming stereotype bias and teachers' judgment accuracy look promising (Pit-ten Cate et al, 2012). However longitudinal studies are necessary to evaluate to what extent the qualitative changes in teachers' judgments resulting from changes in cognitive processes are maintained over time.

In conclusion, to increase judgment accurarcy, one should consider different strategies, including both training of diagnostic competence and focussing on underlying cognitive processes. More specifically, teachers' professional competence should encompass not only teaching knowledge and skills, but also the ability to judge fairly, to assure the validity of learning outcomes. Especially in combination, whereby insights from both education and social psychology could mutually support each other, such strategies can enhance the validity of judgments, which could contribute to a more equitable educational system.

REFERENCES/BIBLIOGRAPHY

Alpert, B., & Bechar, S. (2008). School organizational efforts in search for alternatives to ability grouping. *Teacher and Teacher Education, 24,* 1599–1612. doi:10.1016/j.tate.2008.02.023

Aegisdóttir et al. (2006). The meta-analysis of Clinical Judgment Project: Fifty-six years of accumulated research on clinical versus statistical prediction. *The Counseling Psychologist*, *34*, 341–382. doi:10.1177/0011000005285875

Andrews, T. J., Wisniewski, J. J., & Mulick, J. A. (1997). Variables influencing teachers' decisions to refer children for school psychological assessment services. *Psychology in the Schools, 34*, 239–244. doi:10.1002/(SICI)1520-6807(199707)34:3<239::AID-PITS6>3.0.CO;2-J

Aronson, J., Lustina, M. J., Good, C., Keough, K., Steele, C. M., & Brown, J. (1999). When White men can't do math: Necessary and sufficient factors in stereotype threat. *Journal of Experimental Social Psychology*, *35*, 29–46. doi:10.1006/jesp.1998.1371

Bankowitz, R. A., McNeil, M. A., Challinor, S. M., & Miller, R. A. (1989). Effect of a computer assisted general medicine diagnostic consultation service on house staff diagnostic strategy. *Methods of Information in Medicine*, *28*, 352–356.

Bassili, J. N., & Brown, R. D. (2005). Implicit and explicit attitudes: Research, challenges, and theory. In D. Albarracín, B. T. Johnson, & M. P. Zanna (Eds.), *Handbook of attitudes and attitude change* (pp. 543–574). Mahwah, NJ: Erlbaum.

Berner, E. S., Webster, G. D., Shugerman, A. A., Jackson, J. R., Algina, J., & Baker, A. L. (1994). Performance of four computer based diagnostics systems. *The New England Journal of Medicine*, *330*, 1792–1796. doi:10.1056/NEJM199406233302506

Bodenhausen, G. V., Macrae, C. N., & Sherman, J. W. (1999). On the dialectics of discrimination: Dual processes in social stereotyping. In S. Chaiken, & Y. Trope (Eds.), *Dual process theories in social psychology* (pp. 271–290). New York: Guildford.

Boudreau, L. A., Baron, R. M., & Oliver, P. V. (1992). Effects of expected communication target expertise and timing of set on trait use in person description. *Personality and Social Psychology Bulletin*, *18*, 447–451. doi:10.1177/0146167292184008

Brehmer, A., & Brehmer, B. (1988). What has been learned about human judgment from thirty years of policy capturing? In B. Brehmer & C. R. B. Joyce (Eds.), *Human judgment: The SJT view* (pp. 75–114). Amsterdam: Elsevier.

Brehmer, B. (1994). The psychology of linear judgment models. *Acta Psychologica*, *87*, 137–154. doi:10.1016/0001-6918(94)90048-5

Brophy, J., & Good, T. (1974). *Teacher-student relationships: Causes and consequences*. New York: Holt, Rinehart, and Winston.

Dana, J., & Thomas, R. (2006). In defense of clinical judgment... and mechanical prediction. *Journal of Behavioral Decision Making*, *19*, 413–428. doi:10.1002/bdm

Dawes, R. M. (2002). The ethics of using or not using statistical prediction rules in psychological practice and related consulting activities. *Philosophy of Science*, *69*, S178–S184. doi:10.1086/341844

Dawes, R. M., & Corrigan, B. (1974). Linear Models in Decision Making. *Psychological Bulletin*, *81*, 95–106. doi:10.1037/h0037613

De Houwer, J. (2006). What are implicit measures and why are we using them? In R. W. Wiers & A. W. Stacy (Eds.), *The handbook of implicit cognition and addiction* (pp. 11–28). Thousand Oaks, CA: Sage Publisher.

Enea-Drapeau, C., Carlier, M., & Huguet, P. (2012). Tracking subtle stereotypes of children with trisomy 21: From facial-feature-based to implicit stereotyping. *PLoS ONE*, *7*, e34369. doi:10.1371/journal.pone.0034369

Fazio, R. H. (1990). Multiple processes by which attitudes guide behavior: The MODE model as an integrative framework. In M. P. Zanna (Ed.), *Advances in Experimental Social Psychology* (pp. 75–109, Vol. 23). New York, NY: Academic Press.

Fazio, R. H. (2007). Attitudes as object-evaluation associations of varying strength. *Social Cognition*, *25*, 664–703. doi:10.1521/soco.2007.25.5.603

Fazio, R. H., & Towles-Schwen, T. (1999). The MODE model of attitude-behavior processes. In S. Chaiken & Y. Trope (Eds.), *Dual process theories in social psychology* (pp. 97–116). New York, NY: Guilford Press.

Ferguson, R. F. (2003). Teachers' perceptions and expectations and the Black-White test score gap. *Urban Education*, *38*, 460–507. doi:10.1177/0042085903254970

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. In M. P. Zanna (Ed.), *Advances in Experimental Social Psychology* (Vol. 23, pp. 1–74). New York: Academic Press.

Fiske, S. T., & Taylor, S. E. (1991). *Social cognition* (2nd ed.). New York: McGraw-Hill.

Garb, H. N. (1997). Race bias, social class bias, and gender bias in clinical judgment. *Clinical Psychology: Science and Practice*, *4*, 99–120. doi:10.1111/j.1468-2850.1997.tb00104.x

Gawronski, B., & Bodenhausen, G. V. (2006). Associative and propositional processes in evaluation: An integrative review of implicit and explicit attitude change. *Psychological Bulletin*, *132*, 692–731. doi:10.1037/0033-2909.132.5.692

Getty, D. J., Seltzer, S. E., Tempany, C. M. C., Pickett, R. M., Swets, J. A., & McNeill, B. J. (1997). Prostate cancer: Relative effects of demographic, clinical, histology, and MR imaging variables on the accuracy of staging. *Radiology*, *204*, 471–479.

Gigerenzer, G., & Todd, P. M. (1999). Fast and frugal heuristics: The adaptive toolbox. In G. Gigerenzer, P. M. Todd, & the ABC Research Group (Eds.), *Simple heuristics that make us smart* (pp. 3–34). Oxford: Oxford University Press.

Gilbert, D. T., & Hixon, J. G. (1991). The trouble of thinking: Activation and application of stereotypic beliefs. *Journal of Personality and Social Psychology*, *60*, 509–517. doi:10.1037/0022-3514.60.4.509

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

Glock, S., Klapproth, F., Böhmer, M., & Krolak-Schwerdt, S. (2012). Accountability as a moderator of teachers' tracking decisions: Two experimental studies. In C. A. Shoniregun & G. A. Akmayeva (Eds.), *Ireland International Conference on Education – IICE 2012 proceedings* (pp. 238–243). Basildon, UK: Infonomics Society.

Glock, S., & Krolak-Schwerdt, S. (2013). Does nationality matter? The impact of stereotypical expectations on student teachers' judgments. *Social Psychology of Education*, *16*, 111–127. doi:10.1007/s11218-012-9197-z

Glock, S., Krolak-Schwerdt, S., Klapproth, F., & Böhmer, M. (2012). Improving teachers' judgments: Accountability affects teachers' tracking decisions. *International Journal of Technology and Inclusive Education*, *1*, 89–98.

Goldenberg, C. (1992). The limits of expectation: A case for case knowledge about teacher expectancy effects. *American Educational Research Journal, 29*, 517–544. doi:10.3102/00028312029003517

Gollwitzer, P. M., & Moskowitz, G. B. (1996). Goal effects on action and cognition. In A. W. Kruglanski & E. T. Higgins (Eds.), *Social Psychology: Handbook of basic principles* (pp. 361–399). New York: Guilford Press.

Good, T. L., & Nichols, S. L. (2001). Expectancy effects in the classroom: a special focus opn improving the reading performance of minority students in first-grade classrooms. *Educational Psychologist, 36*, 113–126. doi:10.1207/S15326985EP3602_6

Grove, W. M., Zald, D. H., Lebow, B. S., Snitz, B. E., & Nelson, C. (2000). Clinical versus mechanical prediction: A meta-analysis. *Psychological Assessment*, *12*, 19–30. doi:10.1037/1040-3590.12.1.19

Hattrup, K., & Ford, J. K. (1995). The role of information characteristics and accountability in moderating stereotype-driven processes during social decision making. *Organizational Behavior and Human Decision Proc*, *63*, 73–86. doi:10.1006/obhd.1995.1063

Helmke, A., Hosenfeld, I., & Schrader, F. W. (2004). Vergleichsarbeiten als werkzeug für die verbesserung der diagnostischen Kompetenz von Lehrkräften [Comparative class tests as a tool for improving the diagnostic competence of teachers]. In R. Arnold & C. Griese (Hrsg.), *Schulleitung und Schulentwicklung: Voraussetzungen, Bedingungen, Erfahrungen [School management and school development: Requirements, conditions, experiences]* (S. 119–144). Hohengehren: Schneider.

Hoffrage, U., & Reimer, T. (2004). Models of bounded rationality: The approach of fast and frugal heuristics. *Management Revue, 15*, 437–459.

Hofmann, W., Gschwendner, T., Castelli, L., & Schmitt, M. (2008). Implicit and explicit attitudes and interracial interaction: The moderating role of situationally available control resources. *Group Processes & Intergroup Relations*, *11*, 69–87. doi:10.1177/1368430207084847

Hogarth, R. M., & Karelaia, N. (2007). Heuristic and linear models of judgment: Matching rules and environments. *Psychological Review*, *114*, 733–758. doi:10.1037/0033-295X.114.3.733

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

Hornstra, L., Denessen, E., Bakker, J., van den Bergh, L., & Voeten, M. (2010). Effects on teacher expectations and the academic achievement of students with Dyslexia. *Journal of Learning Disabilities*, *43*, 515–529. doi:10.1177/0022219409355479

Houston, D. A., & Fazio, R. H. (1989). Biased processing as a function of attitude accessibility: Making objective judgments subjectively. *Social Cognition*, *7*, 51–66. doi:10.1521/soco.1989.7.1.51

Johnson, V. E., & Kaplan, S. E. (1991). Experimental evidence on the effects of accountability on auditor judgments. *Auditing: A Journal of Practice & Theory*, *10*, 96–107.

Jussim, L., Eccles, J., & Madon, S. (1996). Social perception, social stereotypes, and teacher expectations: Accuracy and the quest for the powerful self-fulfilling prophecy. A*dvances in Experimental Social Psychology, 28*, 281–388. doi: 10.1016/S0065-2601(08)60240-3

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

Kawakami, K., Dovidio, J. F., Moll, J., Hermsen, S., & Russin, A. (2000). Just say no (to stereotyping): Effects of training in the negation of stereotypic associations on stereotype activation. *Journal of Personality and Social Psychology*, *78*, 871–888. doi:10.1037/0022-3514.78.5.871

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

Krolak-Schwerdt, S., Böhmer, M., & Gräsel, C. (2012). Leistungsbeurteilung von Schulkindern: Welche Rolle spielen Ziele und Expertise der Lehrkraft? [Students' achievement judgments: The role of teachers' goals and expertise]. *Zeitschrift für Entwicklungspsychologie und pädagogische Psychologie*, *44*, 111–122. doi:10.1026/0049-8637/a000062

Kunda, Z., & Spencer, S. J. (2003). When do stereotypes come to mind and when do they color judgment? A goal-based theoretical framework for stereotype activation and application. *Psychological Bulletin*, *129*, 522–544. doi:10.1037/0033-2909.129.4.522

Lee, H., Herr, P. M., Kardes, F. R., & Kim, C. (1999). Motivated search: Effects of choice accountability, issue involvement, and prior knowledge on information acquisition and use. *Journal of Business Research*, *45*, 75–88. doi:10.1016/S0148-2963(98)00067-8

Lerner, J. S., & Tetlock, P. E. (1999). Accounting for effects of accountability. *Psychological Bulletin*, *125*, 255–275. doi:10.1037/0033-2909.125.2.255

Lerner, J. S., & Tetlock, P. E. (2003). Bridging individual, interpersonal, and institutional approaches to judgment and decision making: The impact of accountability on cognitive bias. In S. L. Schneider & J. Shanteau (Eds.), *Emerging perspectives on judgment and decision research* (pp. 431–457). Cambridge: University Press.

Levins, T., Bornholt, L., & Lennon, B. (2005). Teachers' experience, attitudes, feelings and behavioural intentions towards children with special educational needs. *Social Psychology of Education*, *8*, 329–343. doi:10.1007/s11218-005-3020-z

Macrae, C. N., & Bodenhausen, G. V. (2000). Social cognition: Thinking categorically about others. *Annual Review of Psychology*, *51*, 93–120. doi:10.1146/annurev.psych.51.1.93

McCombs, R. C., & Gay, J. (1988). Effects of race, class, and IQ information on judgments of parochial grade school teachers. *The Journal of Social Psychology*, *128*, 647–652. doi:10.1080/00224545.198 8.9922918

McKown, C., & Weinstein, R. S. (2002). Modeling the role of child ethnicity and gender in children's differential response to teacher expectations. *Journal of Applied Social Psychology*, *32*, 159–184. doi:10.1111/j.1559-1816.2002.tb01425.x

Neal, L. V., McCay, A. D., Webb-Johnson, G., & Bridget, S. T. (2003). The effects of African American movement styles on teachers' perceptions and reactions. *The Journal of Special Education*, *37*, 49–57. doi:10.1177/00224669030370010501

Olson, M. A., & Fazio, R. H. (2009). Implicit and explicit measures of attitudes: The perspective of the MODE model. In R. E. Petty, R. H. Fazio, & P. Briñol (Eds.), *Attitudes: Insights from the new implicit measures* (pp. 19–63). New York, NY: Psychology Press.

Orton, R. E. (1996). How can teacher beliefs about student learning be justified? *Curriculum Inquiry*, *26*, 133–146.

Osborne, J. W. (2001). Testing stereotype threat: Does anxiety explain race and sex differences in achievement? *Contemporary Educational Psychology*, *26*, 291–310. doi:10.1006/ceps.2000.1052

Parks, F. R., & Kennedy, J. H. (2007). The impact of race, physical attractiveness, and gender on education majors' and teachers' perceptions of student competence. *Journal of Black Studies*, *37*, 936–943. doi:10.1177/0021934705285955

Pendry, L. F., & Macrae, C. N. (1996). What the disinterested perceiver overlooks: Goal-directed social categorization. *Personality and Social Psychology Bulletin*, *22*, 249–256. doi:10.1177/0146167296223003

Pigott, R. L., & Cowen, E. L. (2000). Teacher race, child race, racial congruence, and teacher ratings of children's school adjustment. *Journal of School Psychology, 38*, 177–195. doi:10.1016/S0022-4405(99)00041-2.

Pit-ten Cate, I. M., Krolak-Schwerdt, S., Glock, S., & Markova, M. (2012). *Orientation decisions concerning the transition from primary to secondary school: The effect of accountability.* Paper presented at the European Conference on Educational Research, Cadiz.

Quinn, A., & Schlenker, B. R. (2002). Can accountability produce independence? Goals as determinants of the impact of accountability on conformity. *Personality and Social Psychology Bulletin*, *28*, 472–483. doi:10.1177/0146167202287005

Reyna, C. (2000). Lazy, dumb, or industrious: When stereotypes convey attribution information in the classroom. *Educational Psychology Review, 12*, 85–110. doi:1040-726X/00/0300-0085

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

Sanbonmatsu, D. M., & Fazio, R. H. (1990). The role of attitudes in memory-based decision making. *Journal of Personality and Social Psychology*, *59*, 614–622. doi:10.1037//0022-3514.59.4.614

Santavirta, N., Solovieva, S., & Theorell, T. (2007). The association between job strain and emotional exhaustion in a cohort of 1.028 Finnish teachers. *British Journal of Educational Psychology*, *77*, 213–228. doi:10.1348/000709905X92045

Scheeler, M. C. (2008). Generalizing effective teaching skills: The missing link in teacher preparation. *Journal of Behavioral Education, 17*, 145–159. doi:10.1007/s10864-007-9051-0

Schuette, R. A., & Fazio, R. H. (1995). Attitude accessibility and motivation as determinants of biased processing: A test of the MODE model. *Personality and Social Psychology Bulletin*, *21*, 704–710. doi:10.1177/0146167295217005

Schwarz, N., & Bohner, G. (2001). The construction of attitudes. In A. Tesser & N. Schwarz (Eds.), *Blackwell handbook of social psychology: Intraindividuel processes* (pp. 436–457). Malden, MA: Blackwell.

Shepard, L. A. (2006). Classroom assessment. In: R.L. Brennan (Ed.), *Educational measurement* (4th ed. pp. 624–646). Westport: Praeger.

Siegel-Jacobs, K., & Yates, J. F. (1996). Effects of procedural and outcome accountability on judgment quality. *Organizational Behavior and Human Decision Processes*, *65*, 1–17. doi:10.1006/obhd.1996.0001

Simonson, I., & Nye, P. (1992). The effect of accountability on susceptibility to decision errors. *Organizational Behavior and Human Decision Processes*, *51*, 416–446. doi:10.1016/0749-5978(92)90020-8

Stangor, C., & McMillan, D. (1992). Memory for expectancy-congruent and expectancy-incongruent information: A review of the social and social developmental literatures. *Psychological Bulletin*, *111*, 42–61. doi:10.1037/0033-2909.111.1.42

Stangor, C., & Schaller, M. (1996). Stereotypes as individual and collective representations. In C. N. Macrae, C. Stangor, & M. Hewstone (Eds.), *Stereotypes and stereotyping* (pp. 3–40). New York: Guilford Press.

Steele, C. M. (1998). A threat in the air: How stereotypes shape intellectual identity and performance. *American Psychologist*, *53*, 680–681. doi:10.1037/0003-066X.53.6.680

Steele, C. M., & Aronson, J. (1995). Stereotype threat and the intellectual test performance of African Americans. *Journal of Personality and Social Psychology*, *69*, 797–811. doi:10.1037/0022-3514.69.5.797

Steele, C. M., Spencer, S. J., & Aronson, J. (2002). Contending with group image: The psychology of stereotype and social identity threat. *Advances in Experimental Social Psychology*, *34*, 379–440. doi:10.1016/S0065-2601(02)80009-0

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

Swanson, B. B. (1986). Teachers judgments of first-graders' reading enthusiasm. *Reading Research and Instruction, 25*(1), 41–46. doi:10.1080/19388078509557857

Swets, J. A., Dawes, R. M., & Monahan, J. (2000). Psychological science can improve diagnostic decisions. *Psychological Science in the Public Interest*, *1*, 1–26. doi:10.1111/1529-1006.001

Tetlock, P. E. (1983). Accountability and complexity of thought. *Journal of Personality and Social Psychology*, *45*, 74–83. doi:10.1037//0022-3514.45.1.74

Tetlock, P. E. (1985). Accountability: A social check on the fundamental attribution error. *Social Psychology Quarterly*, *48*, 227–236.

Tetlock, P. E., & Lerner, J. S. (1999). The social contingency model: Identifying empirical and normative boundary conditions on the error-and-bias portrait of human nature. In S. Chaiken & Y. Trope (Eds.), *Dual process theories in social psychology*. New York: Guilford Press.

Tetlock, P. E., Skitka, L., & Boettger, R. (1989). Social and cognitive strategies for coping with accountability: Conformity, complexity, and bolstering. *Journal of Personality and Social Psychology*, *57*, 632–640. doi:10.1037/0022-3514.57.4.632

Thompson, L. (1995). They saw a negotiation: Partisanship and involvment. *Journal of Personality and Social Psychology*, *68*, 839–853. doi:10.1037/0022-3514.68.5.839

van Dick, R., & Wagner, U. (2001). Stress and strain in teaching: A structural equation approach. *British Journal of Educational Psychology*, *71*, 243–259. doi:10.1348/000709901158505

van den Bergh, L., Denessen, E., Hornstra, L., Voeten, M., & Holland, R. W. (2010). The implicit prejudiced attitudes of teachers: Relations to teacher expectations and the ethnic achievement gap. *American Educational Research Journal*, *47*, 497–527. doi:10.3102/0002831209353594

Wahl,, D., Weinert, F. E., & Huber, G. L. (2007). *Psychologie für die Schulpraxis [Psychology in school]* (2nd revised edition). Belm-Vehrte: Sozio-Publishing.

Weiner, B. (2000). Intrapersonal and interpersonal theories of motivation from an attributional perspective. *Educational Psychology Review, 12*(1), 1–14. doi:10.1023/A:1009017532121

AFFILIATIONS

Ineke Pit-ten Cate
University of Luxembourg

Sabine Krolak-Schwerdt
University of Luxembourg

Sabine Glock
University of Luxembourg

Maria Markova
University of Luxembourg

This study was supported by a grant from the Luxembourgish Fonds National de la Recherche (Grant C10/LM/784116/TRANSINTER/KROLAK_SCHWERDT).

MONIKA TRITTEL, MARA GERICH & BERNHARD SCHMITZ

TRAINING PROSPECTIVE TEACHERS IN EDUCATIONAL DIAGNOSTICS

INTRODUCTION

In their professional routines, teachers have to perform highly complex and demanding tasks (Brante, 2009; Pransky, 2008). They devise and implement learner-centered instruction, manage their classes, promote students' academic achievements, and interact with students, parents, and colleagues. As teachers are challenged by having to deal with a heterogeneous array of academic abilities, interests, and motivations and thus to adapt their educational activities to the individual needs of their students (Vogt & Rogalla, 2009), another key aspect of teachers' professional competences is to be able to accurately diagnose students' abilities and learning behavior.

In prior empirical research, several authors defined the diagnostic competence of teachers as the ability to adequately judge students' characteristics and the requirements of the tasks that the students are required to perform (e.g., Artelt & Gräsel, 2009; Schrader, 2006). This approach of defining teachers' diagnostic competence primarily focuses on the accurate diagnosis of students' achievements. Exceeding this definition and referring to the cross-curricular diagnosis of learning behavior, Klug (2011) defines the diagnostic competence of teachers as the "ability to interpret students' academic growth and their growth in using learning strategies" (p. 12). Considering several studies (e.g., McDermott & Breitman, 1984; Yen, Konold, & McDermott, 2004) that have supported the connection between students' learning behavior and their academic achievement, this definition points out the importance of adequately diagnosing learning behavior as a central precondition of being able to foster the individual academic growth of students. The aim of the diagnosis should be to identify information that will allow for specific pedagogical decisions and actions (Carpenter, Fennema, Peterson, & Carey, 1988; Helmke, Hosenfeld, & Schrader, 2004; Vogt & Rogalla, 2009).

Based on this approach, Klug, Bruder, Kelava, Spiel, and Schmitz (2013) developed and empirically tested a model of teachers' diagnostic competence that represents the process of diagnosing learning behavior. Following the terminology used in process models of self-regulation (e.g., Schmitz & Wiese, 2006), it describes the diagnosis of learning behavior as a three-dimensional process that consists of a preactional, an actional, and a postactional phase. As can be seen in Figure 1, the model has a cyclical character. The three phases are ordered in time and influence each other.

S. Krolak-Schwerdt et al., (Eds.), Teachers' Professional Development, 63–78.

The preactional phase (first dimension) contains every action of the diagnosis before summing the information to obtain an actual diagnosis of a student's learning behavior. The teacher has to set the aim of the diagnosis to be able to monitor the individual student's learning process and to provide support to the student based on the diagnosis. Furthermore, the basic diagnostic skills that the teacher possesses are activated in the preactional phase. These basic diagnostic skills comprise knowledge about methods for gathering information, assessing the psychological quality criteria of tests, and making judgments.

The actional phase (second dimension) includes skills that are important when the actual diagnostic action takes place. The teacher should act systematically to make a reliable diagnosis. The systematic action should begin with making a prediction about a student's development and any possible underlying learning difficulties. After that, the teacher should gather information from different sources and choose the relevant ones to finally interpret the data and come to a final diagnosis.

In the postactional phase (third dimension), which begins immediately after a diagnosis has been made, pedagogical action that follows from the diagnosis should be implemented. First, giving feedback to the student and to the student's parents is a key component of subsequent pedagogical action. Furthermore, writing down plans for the individual student's advancement is another content area of pedagogical action that should be implemented after the diagnosis. Finally, adapting the class in reaction to the diagnosis by means of teaching appropriate learning strategies and self-regulated learning is relevant for supporting students in their academic development.

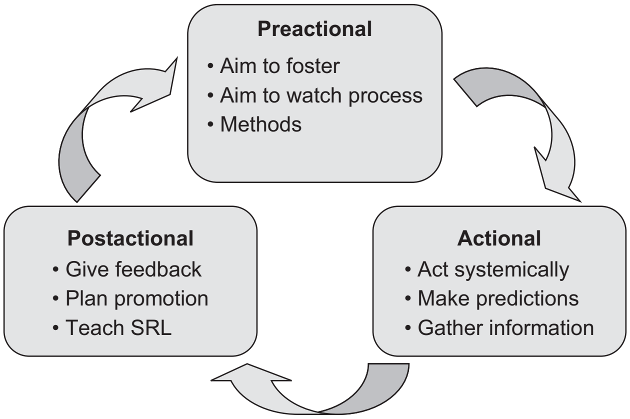

Figure 1.Process model of teachers' diagnostic competence concerning students' learning behavior (Klug et al., 2013).

To be able to meet these complex and diverse requirements in their daily diagnoses, prospective teachers should be trained in professional diagnostic competences in their university and postgraduate studies. In this context, it is essential to particularly focus on specific instructional practices that enhance students' practical diagnostic competences in addition to teaching mere theoretical knowledge (Döbrich, Klemm, Knauss, & Lange, 2003; Halász, Santiago, Ekholm, Matthews, & McKenzie, 2004; Oser, 2001; Terhart, 2000).

In spite of this well-known necessity of assisting prospective teachers in their professional development, we were able to show in a pilot study that, until now, professionalization in diagnostics has not been well implemented in German university teacher training programs (the first phase of teacher education in Germany) and rather focuses on the generation of theoretical knowledge instead of the development of practical competences. In this study, we asked a broad sample of n=99 expert grammar school teachers about their university education in educational diagnostics. Results revealed that only 15% of the teachers surveyed had taken part in courses on educational diagnostics in their university education and that these courses scarcely included the practical application of the theoretical diagnostic knowledge that they gained.

Moreover, connections between this theoretical knowledge and future practice are often not made explicit (Bransford, Brown, & Cocking, 2000; Grossman, 2005). As a consequence, prospective teachers do not feel well prepared for their profession (Herzog & von Felten, 2001) and show deficits in essential pedagogical competences, particularly in diagnostic competences (Klug et al., 2013; Spinath, 2005). Finally, these deficits result in the so-called reality shock that frequently confronts new teachers when they enter their careers and take on the responsibilities of their roles as schoolteachers (Stokking, Leenders, De Jong, & Van Tartwijk, 2003; Veenman, 1984).

Based on these findings, there is a growing request for educational programs to foster teachers' diagnostic competence (Klieme et al., 2003). As numerous studies on the effectiveness of different approaches and methods in the context of the development of practical competences have revealed, such programs should involve a lot of active learning (Garet, Porter, Desimone, & Yoon, 2001; Grossman, 2005) and work on specific case studies (Kolodner, Gray, & Fasse, 2003).

Furthermore, teaching students to reflect on their own professional practice seems to be another key component of effective teacher training, as reflection is assumed to be essential for professional development (Boud, 1999; Calderhead & Gates, 1993; Wilson & Berne, 1999) and ensures the transfer of theoretically acquired knowledge to a teacher's future professional routine (Eraut, 2003). In addition to competence-based training designs, the evaluation of the effectiveness of teacher training programs should be based on competence-oriented assessment approaches, too (Kunter, 2011). For this, adequate instruments are needed.

To meet the outlined need for measures in order to enhance teachers' diagnostic competence, the purpose of the present study was to develop and evaluate a training program in educational diagnostics for prospective teachers with a specific focus on the development of practical competences that are actually relevant to teachers' future professional work.

TRAINING PROGRAM ON EDUCATIONAL DIAGNOSTICS FOR PROSPECTIVE TEACHERS

In order to foster prospective teachers' diagnostic competence, we developed and implemented a hands-on seminar on educational diagnostics for undergraduates who are on their way to becoming grammar school teachers (the first phase of teacher training in Germany). The seminar focuses on formative rather than summative diagnostics.

It comprises weekly units of 100 minutes each over a period of ten weeks. Each unit is followed by homework in which the undergraduates respond to knowledge questions, elaborate on their own diagnostic case scenario, and reflect on their individual goals for the seminar. In Table 1, the sequence of units is displayed as well as the units' affiliation to the model of diagnostic competence by Klug and colleagues (2013), which serves as scaffolding for the training program. The units' contents are described in this regard in more detail in the following paragraphs.

General Outline for the Configuration of the Training Program

The undergraduates who participate in the training program on educational diagnostics find themselves in a situation in which they are confronted with a practical issue of high importance for their future work, but they have hardly any practice in teaching. They lack experience from school in the role of a teacher, but they are supposed to develop a professional competence that, by definition, goes far beyond theoretical knowledge. Consequently, the training program has to compensate for this missing experience and find a way to connect the contents of the seminar to real-life situations and problems. Therefore, the key method used in each of the units is to work with self-generated diagnostically relevant cases. Participants were initially encouraged to describe a case that they know from their own time in school or from the reports of others. From that point on, they constantly work on this case in the weekly homework and describe each step in the diagnostic process from the goal of the diagnosis to their educational reaction.

In addition to that, the training units themselves are configured explicitly with regard to the application of themes. The participants engage in discussions and simulations that are meant to enable them to anticipate the benefits as well as the difficulties of educational diagnostics in their later work as teachers.

Generally, the participants in the seminar do not just passively receive the components of the diagnostic process, but rather work on them actively. For example, they develop diagnostic instruments, practice providing feedback, and plan palpable supportive measures, which they think about and discuss in chaired discussions. Thereby, the students receive an impression of peer-consulting as well.

Table 1. Overview of the seminar on educational diagnostics

Diagnostic Phase	*Unit*	*Title*
Pre-actional	1	Introduction to Educational Diagnostics
	2	Quality Criteria of Diagnostic Judgment and Prevention of Judgment Errors
	3	Diagnostic Methods and Instruments
Actional	4	Evaluation of Diagnostic Information
	5	Interpretation and Feedback of Diagnostic Results
Overall	6	Scope of Application I: Diagnoses of Performance
	7	Scope of Application II: Diagnoses of Learning
	8	Scope of Application III: Diagnoses of Tuition
Post-actional	9	Educational Reaction I: Support Cycle and Student Support Plans
	10	Educational Reaction II: Supportive Measures

Pre-Actional Phase of Diagnosing: Units 1-3

The first three units of the diagnostic training program deal with the pre-actional phase of diagnosing. The focus lies on preparing to take actual diagnostic action. In unit 1, participants become acquainted with the definition, the goals, and the scope of educational diagnostics for teachers. They become familiar with several arguments for engaging in educational diagnostics, for instance, teachers' official assignment to deliver educational judgments and empirical evidence for the benefits of diagnosing appropriately and efficiently. In unit 2, the participants learn to transfer the basic quality criteria of diagnostic judgments – objectivity, reliability, and validity – to the context of school and educational judgments. By means of specific case examples, they also experience typical judgment errors such as the Halo-Effect and the Fundamental Attribution Error, which easily occur in specific situations and when judges are not educated appropriately. As ways to prevent these errors, cooperation with colleagues and the identification of individual judgment tendencies are discussed. In unit 3, the participants are taught further predisposition for being able to adequately plan diagnostic actions: They become familiar with diverse diagnostic methods, the distinction of instruments, and implications for or against the detailed methods. At the end of this unit, the participants should be able to select an appropriate and efficient diagnostic method and a corresponding instrument to collect diagnostic information and, finally, to respond to a specific diagnostic question.

Actional Phase of Diagnosing: Units 4-5

In the actional phase of diagnosing, pieces of diagnostic information are collected using the method(s) and instrument(s) selected at the end of the pre-actional phase.

The next steps in the diagnostic process are the evaluation of diagnostic information and the interpretation of diagnostic results. The evaluation is taught in unit 4. The diverse methods of analysis as well as their conditions are presented (e.g., the level of data required for the statistical analyses). Through an illustrative exercise, participants discover how different kinds of underlying data can lead to the same mean value. To do this illustration, some students who represent test results line up in front of the group such that the students standing at the one end represent very high results, whereas students standing at the other end represent very low results. One student will be positioned within the line to represent the statistical mean of the results. With various arrangements of the results, the respective value of the mean can be shown. In particular, this illustration can be used to show that a mean close to the middle of the range of values can be created by averaging single values that are close to that mean value, but that alternatively, the same mean value can also be constituted by some high and some low single values without any single value being close to the mean value.

Unit 5 is concerned with the interpretation and feedback of the diagnostic results. Diverse reference standards are compared to each other regarding their formative effect on students, pointing out the motivating function of the individual reference standard. Furthermore, participants collect different ways to give feedback on diagnostic results; for example, by using brief written comments, charts and tables, and feedback discussions with students. Feedback is demonstrated in a scripted simulation in which one participant acts as a teacher who gives feedback and another participant acts as a student who receives the feedback. The other participants of the seminar observe and analyze the feedback talk, point out the positive aspects, and suggest improvements. Both actors reflect on their feelings in the role of the deliverer of feedback versus the receiver and make suggestions to create an agreeable atmosphere for everyone involved.

Scope of Application: Units 6-8

Units 6-8 focus, respectively, on one of the diverse topics that the educational diagnoses can be applied to. Hence, these units are not related to one of the diagnostic phases but rather illustrate applications in the fields of performance diagnoses, learning diagnoses, and tuition diagnoses. The participants become acquainted with the variation in diagnostic goals, methods, and instruments according to the topic.

The most prominent topic or aim of diagnostics in school is the diagnosis of students' performance. Teachers have to judge performances in the various topics they teach, they have to allocate justified and objective marks, and they must develop adequate instruments to map these performances appropriately. This assignment of teachers is taught in unit 6. Participants discuss different approaches for defining performance and reflect upon predispositions for academic performance such as motivation, learning behavior, and the general set-up of testing. The prospective teachers discover how strong of an influence they have on these variable aspects and learn about formative rather than summative diagnoses of performance.

Unit 7 uncovers the opportunities and methods of diagnosing learning behavior. These can be regarded as important predecessors of performance and as providing great opportunities for teachers to support students individually. The participants use observation sheets to structure their observations during lessons and to keep records of them. They develop the ability to turn incidental subjective observations into professional diagnostic observations. They learn to create optimal settings for observations (and to understand the necessity of observation-free periods!). Promoting students' self-assessments of learning via learning diaries or partner diagnoses is picked out as a central theme, too. One key insight in this unit is the following: The same observable indicators of learning behavior (e.g., a student looking out of the window during a lesson) can occur for dramatically different reasons (e.g., the student is lost in thought versus the student is reflecting intensely on a class topic). It is often necessary to consider observed indicators from different perspectives, thus supporting the need for dialogic diagnoses, which means that the student's perspective should be taken into consideration to come to an adequate interpretation of diagnostic information.

The third application topic addressed in unit 8 of the seminar concerns the diagnosis of tuition (i.e., teaching). Participants are shown methods and instruments that can be used to judge the quality of their lessons and the effects of their tuition on individual learning. Thus, the diagnostic focus in this context lies more on the teacher than on the students. Diagnoses of tuition help teachers to further develop the instruction they give and to allocate their efforts appropriately, i.e., effectively and efficiently. Once more, collegial cooperation is recommended as a means to perform tuition diagnoses and development.

Post-Actional Phase of Diagnosing: Units 9-10

Following the phases in the diagnostic process that are worked out before, units 9 and 10 enable the participants to take the final and perhaps most crucial step: the educational reaction to diagnoses. Educational diagnostics are not performed for their own purpose. Ideally, the diagnostic process aims to foster students individually (formative assessment). Coming to a professional accurate diagnosis is important, but rather ineffective if subsequent educational actions are missing. If the diagnostic process is performed professionally, it is possible to directly link the diagnosis to supportive measures for students: Formative diagnoses comprise a student's deficit or problem (as a task that needs to be handled) and his or her strengths (as a point to start from and to build upon).

In unit 9, the cyclical nature of student support is pointed out; for example, the necessity of evaluating the effect of the support. In Germany, the so-called student support plan is used as a tool to record and to plan supportive measures; these plans are compulsory in certain cases such as the repetition of a grade. Participants learn to fix central elements of a support plan: diagnosis, conjoint agreement on goals, supportive measures, and duration.

Supportive measures, for the enhancement of self-regulated learning in particular, are developed in unit 10. As a particularly fruitful and efficient educational reaction, the power of feedback is presented and discussed.

HOW PROSPECTIVE TEACHERS BENEFIT FROM PRACTICAL DIAGNOSTIC TRAINING

The training program on diagnostic competence for prospective teachers is believed to foster the theoretical and practical diagnostic skills of its participants. To ensure the effectiveness and, consequently, the efficiency of this extensive course, a respective evaluation is required. Because we are targeting the applied diagnostic skills of the participants in particular, the evaluation method has to be able to determine participants' abilities in order to put the training contents into action.

Because we presented a detailed description of the training program on diagnostic competence for prospective teachers in the previous section of this chapter, the description will not be repeated in the following account of the method that we applied in the current study. As the collection of data is still ongoing, we will present preliminary results based on an initial fraction of the sample.

Participants

The current sample consists of n=34 prospective teachers who are in the process of obtaining the first degree in Education with the intention of becoming grammar school teachers. At the time in which we conducted this study, they were studying at a German university in the degree program of the first phase of teacher education in Germany. The participants were on average 24.91 years old with a standard deviation of 4.84 years. A total of 50 % were male. They participated voluntarily in the study.

Design

Using a quasi-experimental within-subjects design, we aimed to compare the diagnostic competence of the training group with a control group whose participants were engaged in another course for prospective teachers but not in the diagnostic training program. The undergraduates chose several compulsory courses from the list of available courses. Those who decided to take the diagnostic course and agreed to participate in the study constituted the experimental group. Those who decided to take any other course and agreed to participate in the study constituted the control group. The courses, Diagnostics versus Alternative, served as the factor levels of the independent variable "Intervention". The dependent variables and their measurement will be described in the next section.

Table 2 shows the design of the evaluation study. So far, the experimental group has already completed the study. Thus, the course "Educational Diagnostics for Prospective Teachers" has taken place and the data collection for this group is complete. At the moment, the control group consists of only a fraction of the scheduled sample; the data collection of another fraction of the control group is still occurring.

Table 2. Design of the evaluation study

Group	*Pre-Test*	*Intervention*	*Post-Test*	*Follow-Up Test*
Control	x	Alternative	x	-
Experimental	x	Educational Diagnostics	x	x

Procedure

All study participants completed a paper-and-pencil test to measure their individual diagnostic competence at the beginning of the selected course (Diagnostics versus Alternative course), at the end of the course, and again approximately 4 weeks afterwards. The course on educational diagnostics consisted of ten units. These units were already described in detail. The data were collected anonymously by using a unique individual code for each participant. The experimental group received individual feedback in written form concerning their diagnostic competence, knowledge, and professional self-concept after the pre-test and after the post-test.

Instruments and Measures

The diagnostic competence and several predictors were measured using the model of diagnostic competence of teachers and the respective instrument developed and evaluated by Klug and colleagues (2013). This paper-and-pencil test comprises a case scenario with open-ended questions to measure diagnostic competence, a multiple-choice test to assess diagnostic knowledge, and a self-assessment questionnaire to measure the person's professional self-concept and reflected diagnostic experience.

For the purpose of uncovering applied diagnostic skills (and, as a prerequisite, applied diagnostic knowledge), the diagnostic knowledge test was modified in parts. These modifications were made to show selected diagnostic skills in exertion and were intended to minimize the occurrence of correct solutions by chance. For some items, the response format was changed from multiple-choice to open-ended questions, and new questions were introduced. An overview of these modifications has been assembled in Table 3.

Table 3. Modifications of the diagnostic knowledge test regarding applied knowledge

Original item	*Modified item*	*Original response format*	*Modified response format*
Which diagnostic methods do you know?	Specify five different diagnostic methods.	Multiple-choice	Open-ended
How do you structure a counseling interview with parents to report learning difficulties?	Specify three different methods that can be used to report diagnostic results.	Multiple-choice	Open-ended
Nonexistent	A task with 0-3 points is executed with 0 points obtained by half of a class containing 20 total students. The other students in this class receive 3 points each. Compute the task difficulty.	-	Open-ended
Nonexistent	Formulate a suitable item with an adequate response format… … for an observation sheet	-	Open-ended
Nonexistent	Formulate a suitable item with an adequate response format… … for a learning diary	-	Open-ended

Results

For the purpose of evaluating the practical seminar on educational diagnostics for prospective teachers, several analyses were computed using the items on applied diagnostic knowledge as well as the overall measure of diagnostic competence.

Effects of the intervention on diagnostic competence and on applied knowledge. To evaluate the effect of the training program on the diagnostic competence of the participants, an analysis of variance with repeated measures was calculated. The results of the case scenario measuring diagnostic competence on the pre-test and post-test were compared between the experimental and control groups. Statistics showed a significant interaction of treatment and time: There was a positive development in diagnostic competence for the experimental group but not for the control group ($F(1,32) = 43.28$, $p = .00$, $\eta_p^2 = .58$). Figure 2 illustrates this finding.

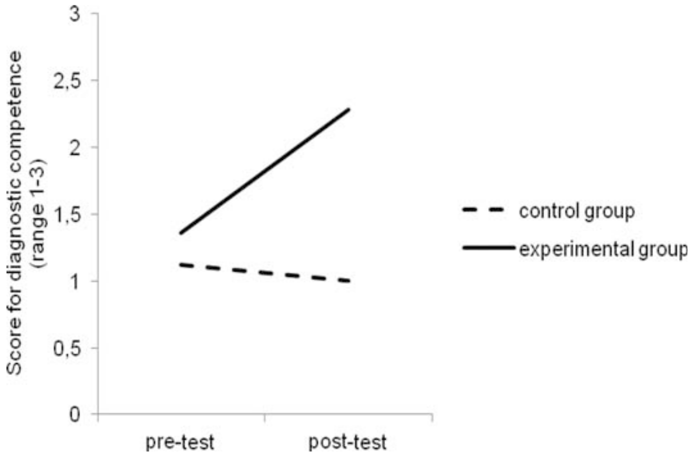

Figure 2. Effect of the intervention on diagnostic competence.

To evaluate the effect of the training program on the applied diagnostic knowledge of the participants, a doubly multivariate analysis of variance was computed: The results of the five applied knowledge items on the pre-test and post-test were compared between the experimental and control groups. Table 4 contains the respective results. Because of the statistical significance of the time by group interaction, only these results are reported here.

Connection of applied diagnostic knowledge and diagnostic competence. To determine the connections between the items that were designed to assess applied knowledge and the measure of diagnostic competence – an applied construct by nature – a correlation was computed between the applied knowledge items as a scale and diagnostic competence. One-tailed significance testing was applied in this analysis because of the assumption that high applied knowledge would go along with high diagnostic competence.

The aggregation of the five applied knowledge items into an “applied knowledge” scale seems justified because of relatively high internal consistencies (Cronbach’s alpha) on the pre-test (.75) and post-test (.91).

Table 4.Synopsis of the results of a doubly multivariate ANOVA for applied diagnostic knowledge

				Time x Group interaction		
	Applied Knowledge Item	*Pre-Test Mean (SD)*	*Post-Test Mean (SD)*	$F_{(1,32)}$	*p*	η_p^2
EG CG	Specify five different diagnostic methods.	1.87 (1.26) 0.10 (0.32)	4.17 (1.66) 0.00 (0.00)	13.50	.001	.30
EG CG	Specify three different methods that can be used to report diagnostic results.	0.10 (0.32) 1.21 (0.98)	2.79 (0.51) 0.20 (0.63)	18.60	.000	.37
EG CG	[…]Compute the task difficulty.	0.54 (1.10) 0.40 (0.70)	2.29 (1.16) 0.20 (0.63)	13.83	.001	.30
EG CG	Formulate a suitable item with an adequate response format for an observation sheet.	0.67 (1.10) 0.00 (0.00)	2.17 (0.96) 0.00 (0.00)	12.10	.001	.27
EG CG	Formulate a suitable item with an adequate response format for a learning diary.	0.38 (0.77) 0.00 (0.00)	2.42 (1.02) 0.00 (0.00)	21.92	.000	.41

Note. Experimental Group EG (n=24); Control Group CG (n=10).

The aggregation of the five applied knowledge items into an „applied knowledge" scale seems justified because of relatively high internal consistencies (Cronbach's alpha) on the pre-test (.75) and post-test (.91).

A one-tailed bivariate correlation of the scales "applied knowledge" and "diagnostic competence" on the post-test revealed a statistically significant value of *r*(32)= .84, *p*<.001.

DISCUSSION

The aim of the present study was to develop and investigate the effects of a training program in educational diagnostics for prospective teachers with a specific focus on practical diagnostic competences that have actual relevance for future professional work. Consequently, it was hypothesized that participation in the training program would enhance participants' practical diagnostic competences in addition to the development of their theoretical knowledge. To test this hypothesis, a competence-based assessment approach was used. As expected, the results revealed positive effects of the training program, which manifested in the experimental group in the application of participants' acquired diagnostic knowledge as well as in their overall diagnostic competence. Consequently, the findings of our study led us to the conclusion that it is possible to enhance prospective teachers' diagnostic competence through training, even if the single components of diagnostic competence are each addressed for only a relatively short time.

By finding a correlation between applied knowledge and diagnostic competence, we were able to fortify our assumption that higher applied diagnostic knowledge goes along with higher diagnostic competence. Future research is needed to investigate whether there is a causal relation.

Although the present study yielded important findings highlighting the effects of the specified training program on prospective teachers' diagnostic competence, there are several limitations that need to be noted and should be addressed in future studies. Because of the diverse characteristics of the sample, it did not turn out to be representative. First, the control group consisted of ten persons only. Hence, a central aim of further investigations should be to recruit a larger sample. Second, because undergraduates were free to choose to take part in the diagnostic course (experimental group) or to engage in another course (control group), they could not be randomly assigned to the experimental conditions. Thus, we cannot rule out that our results were somewhat influenced by self-selection. We have to consider the possibility that the participants who took part in the diagnostic course were the more motivated ones from the population of prospective teachers. To demonstrate the generalizability of our findings to the population of prospective teachers, it will be necessary to determine whether the results can be replicated using a more representative sample as well as random assignment.

Based on the results and limitations of the present study, several recommendations can be made for future research. The specified methods of intervention and assessment proved beneficial in the context of the enhancement of prospective teachers' diagnostic competences, particularly with regard to the development of practical competences. According to these findings, further investigations in the field of other teacher competences may use the supporting effects of active and competence-based training as well. For example, Gerich, Trittel, and Schmitz (2012) developed and evaluated a training program to foster teachers' counseling competence. Analogical to the present study, the program included large sequences of active learning and working on specific case studies from the participants' daily professional routines. Furthermore, another key component of the program required participants to reflect on their own professional practice in order to increase the development of their counseling competences and to ensure the transfer of their acquired skills and knowledge to their actual professional work. The evaluation of the program was also based on competence-oriented assessment strategies, for which adequate instruments - similar to the instruments used in the present study - have been constructed and utilized.

This leads us to another important implication for future research, which consists of the development and validation of competence-based assessment strategies with respect to other professional skills of teachers. Only the application of such assessment approaches will allow for the actual practical competences of teachers to be accurately measured and for a reasonable evaluation of the effectiveness of teacher education programs (Kunter, 2011). The development of new strategies and the continuing advancement of existing instruments prove to be future challenges. For example, the assessment of teachers' diagnostic competences may be advanced by analyses of real work samples or classroom observations to allow measures of teachers' diagnostic competences in their actual professional routine and thereby to permit even more precise measurements than the outlined case scenario and the specified open-ended knowledge items. However, these attempts should simultaneously consider the efficiency of the strategies and instruments that have been developed, particularly with regard to their application to larger sample sizes. Up to now, standardized instruments like scenariotests, competence-oriented knowledge tests, or for instance, situational judgment tests have been the methods of choice because they are able to measure real teacher behavior but are still economical (Hedlund, Witt, Nebel, Ashford, & Sternberg, 2006; Rivard, Missiuna, Hanna, & Wishart, 2007).

Finally, because the diagnosis of learning behaviour has been acknowledged to be an everyday task of teachers at school and particularly because we found a conspicuously low base level of prospective teachers' diagnostic competence in our study, the training of this central teacher competence should receive more attention and become a fixed component of early teacher education. In this context, new curricula of teacher education - not only concerning diagnostic skills but also other essential teacher competences - should be developed and implemented with a specific focus on the improvement of practical competences that go beyond theoretical knowledge.

AUTHOR NOTE

The preparation of this chapter was supported by grant SCHM 1538/5-3 from the German Research Foundation (DFG) in the Priority Programme "Competence Models for Assessing Individual Learning Outcomes and Evaluating Educational Processes" (SPP 1293).

REFERENCES

Artelt, C., & Gräsel, C. (2009). Diagnostische Kompetenz von Lehrkräften – Gasteditorial [Diagnostic competence of teachers - Guesteditorial]. *Zeitschrift für Pädagogische Psychologie[German Journal of Educational Psychology], 23*, 157–160.

Boud, D. J. (1999). *Experience and learning. Reflection at work*. Deakin: Deakin University, Faculty of Education.

Bransford, J. D., Brown, A. L., & Cocking, R. R. (Eds). (2000). *How people learn. Brain, mind experience, and school. Commission on Behavioral and Social Sciencesand Education.* Washington, DC: National Research Council, National Academy Press.

Brante, G. (2009). Multitasking and synchronous work. Complexities in teacher work. *Teaching and Teacher Education, 25*, 430–436.

Calderhead, J., & Gates, P. (1993). *Conceptualizing reflection in teacher development*. London: Falmer Press.

Carpenter, T. P., Fennema, E., Peterson, P. L., & Carey, D. A. (1988). Teachers' pedagogical content knowledge of students' problem solving in elementary arithmetic. *Journal for Research in Mathematics Education, 19*, 385–401.

Döbrich, P., Klemm, K., Knauss, G., & Lange, H. (2003). *Attracting, developing and retaining effective teachers.* Supplement to the country background report for the Federal Republic of Germany.

Eraut, M. (2003). *Learning during the first three years of postgraduate employment - The LiNEA Project.* Paper presented at the 10th biennial conference of the European Association for Research on Learning and Instruction (EARLI). Padova, Italy.

Garet, M. S., Porter, A. C., Desimone, L., Birman, B. F., & Yoon, K. S. (2001). What makes professional development effective? Results from a national sample of teachers. *American Educational Research Journal, 38*, 915–945.

Gerich, M., Trittel, M., & Schmitz, B. (2012). Förderung der Beratungskompetenz von Lehrkräften durch Training, Feedback und Reflexion. Methodenhandlungsorientierter Intervention und Evaluation [Promoting teachers' counseling competence by training, feedback, and reflection. Methods of action-oriented intervention and evaluation]. In M. Kobarg, C. Fischer, I. M. Dalehefte, F. Trepke, & M. Menk (Eds.), *Lehrerprofessionalisierungwissenschaftlichbegleiten – Strategien und Methoden [Scientifical monitoring of teacher-professionalization – Strategies and methods]*. Münster: Waxmann.

Grossman, P. (2005). Research on pedagogical approaches in teacher education. In M. Cochran-Smith & K. M. Zeichner (Eds.), *Studying teacher education. The report of the AERA panel on research and teacher education* (pp. 425–476). Mahwah, NJ: Erlbaum.

Halász, G., Santiago, P., Ekholm, M., Matthews, P., & McKenzie, P. (2004). *Attracting, developing, and retraining effective teachers.* Paris: OECD.

Hedlund, J., Witt, J. M., Nebel, K. L., Ashford, S. J., & Sternberg, R. (2006). Assessing practical intelligence in business school admissions. A supplement to the graduate management admission test. *Learning and Individual Differences, 16*, 101–127.

Helmke, A., Hosenfeld, I., & Schrader, F.-W. (2004). Vergleichsarbeiten als Instrument zur Verbesserung der Diagnosekompetenz von Lehrkräften [Comparative tests as an instrument for the improvement of diagnosic competence of teachers]. In R. Arnold, & C. Griese (Eds.), *Schulmanagement und Schulentwicklung [Schoolmanagement and school development]* (pp. 119–144). Hohengehren: Schneider.

Herzog, W., & von Felten, R. (2001). Erfahrung und Reflexion. Zur Professionalisierung der Praktikumsausbildung von Lehrerinnen und Lehrern [Experience and reflection. At the professionalisation of the practical education of teachers]. *BeiträgezurLehrerbildung [Contributions to teacher education], 19*, 17–28.

Klieme et al. (2003). *Expertise zurEntwicklungnationalerBildungsstandards. [Expertise for the development of national educational standards]*. Berlin: BundesministeriumfürBildung und Forschung (BMBF) [German Federal Ministry of Education and Research].

Klug, J. (2011). *Modeling and training a new concept of teachers' diagnostic competence.* Dissertation, TU, Darmstadt. Retrieved from http://tuprints.ulb.tu-darmstadt.de/2838/

Klug, J., Bruder, S., Kelava, A., Spiel, C., & Schmitz, B. (2013). Diagnostic competence of teachers. A process model that accounts for diagnosing learning behavior tested by means of a case scenario. *Teaching and Teacher Education, 30*, 38–46.

Kolodner, J. L., Gray, J., & Fasse, B. B. (2003). Promoting transfer through case-based reasoning. Rituals and practices in learning by DesignTM classrooms. *Cognitive Science Quarterly, 3*, 183–232.

Kunter, M. (2011). Theorie meets Praxis in der Lehrerbildung – Kommentar [Theory meets practice in teacher education - Commentary]. *Erziehungswissenschaft [German Journal of Educational Science], 22*, 107–112.

McDermott, P. A., & Breitman, B. S. (1984). Standardization of a scale for the study of children's learning styles. Structure, stability, and criterion validity. *Psychology in the Schools, 21*, 5–13.

Oser, F. (2001). Standards: Kompetenzen von Lehrpersonen [Standards: Competences of teachers]. In F. Oser & J. Oelkers (Eds.), *Die Wirksamkeit der Lehrerbildungssysteme. Von der Allrounderausbildung zur Ausbildung professioneller Standards [Effectivity of teacher education systems. From an* *allrounder education to the development of professional standards]*. Nationales Forschungsprogramm 33, Wirksamkeit unserer Bildungssysteme. Zürich: Rüegger.

Pransky, K. (2008). *Beneath the surface. The hidden realities of teaching culturally and linguistically diverse young learners.* Portsmouth: Heinemann.

Rivard, L. M., Missiuna, C., Hanna, S., &Wishart, L. (2007). Understanding teachers` perceptions of the motor difficulties of children with developmental coordination disorders (DCD). *British Journal of Educational Psychology, 77*, 633–648.

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.

Schrader, F.-W. (2006). DiagnostischeKompetenz von Eltern und Lehrern [Diagnostic competence of parents and teachers]. In D. H. Rost (Ed.), *Handwörterbuch Pädagogische Psychologie [Concise dictionary of Educational Psychology]* (pp. 95–100). Weinheim: Beltz.

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 [German Journal of Educational Psychology], 19*, 85–95.

Stokking, K., Leenders, F., De Jong, J., & Van Tartwijk, J. (2003). From student to teacher. Reducing practice shock and early dropout in the teaching profession. *European Journal of Teacher Education, 26*, 329–350.

Terhart, E. (Ed.). (2000). *Perspektiven der Lehrerbildung in Deutschland. Abschlussbericht der von der Kultusministerkonferenz eingesetzten Kommission [Perspectives of teacher education in Germany. Final report of the committee appointed by the conference of ministers for education]*. Weinheim: Beltz.

Veenman, S. (1984). Perceived problems of beginning teachers. *Review of Educational Research, 54*, 143–178.

Vogt, F., & Rogalla, M. (2009). Developing adaptive teaching competency through coaching. *Teaching and Teacher Education, 25*, 1051–1060.

Wilson, S. M., & Berne, J. (1999). Teacher learning and the acquisition of professional knowledge. An examination of research on contemporary professional development. *Review of Research in Education, 24*, 173–209.

Yen, C. J., Konold, T. R., & McDermott, P. A. (2004). Does learning behavior augment cognitive ability as an indicator of academic achievement? *Journal of School Psychology, 42*, 157–169.

AFFILIATIONS

Monika Trittel
Institute of Psychology, Working Group for Educational Psychology,
Technische Universität Darmstadt, Germany

Mara Gerich
Institute of Psychology, Working Group for Educational Psychology,
Technische Universität Darmstadt, Germany

Bernhard Schmitz
Institute of Psychology, Working Group for Educational Psychology
Technische Universität Darmstadt, Germany

JAN D. VERMUNT

TEACHER LEARNING AND PROFESSIONAL DEVELOPMENT

INTRODUCTION

The societal demand for evidence that teacher education and professional development initiatives result in improved student learning outcomes is increasing. In Figure 1, a chain of evidence from teacher education and professional development to student learning outcomes is depicted. Teacher education and professional development programs initiate (student) teachers' learning processes, leading to teachers' learning outcomes. These outcomes may be conceptualized as increased or changed knowledge, understandings, intentions, practices and motives/emotions. When teachers use these knowledge, practices, etc. in their teaching, they form an important element of the learning environment for the students, together with the learning materials, physical environment, fellow-students, etc. Teachers' teaching practices initiate students' learning processes which, in turn, lead to students' learning outcomes in terms of increased or changed knowledge and understanding of the subject matter, skills, motivations, emotions, etc.

Of course, influences can also go the other way around. Teachers may learn a lot from studying the learning processes of their students: the way they learn, how their approaches to learning affect their learning outcomes, how their teaching affects students' learning, etc. And students may learn a lot from studying their teachers' learning, since teachers are supposed to be experts in learning. The quality of the learning outcomes students achieve may give rise to implementing changes in the pedagogy and content of the teacher education program.

There is, however, almost no research that covers the whole chain of causation sketched above. One reason may be that the research domains of student learning and teacher learning are organized in separate research communities, with each their own professional organizations, scientific journals, special interest groups, etc. Another reason may be that covering the whole chain transcends the duration of an average research project.

Traditional boundaries have to be crossed to achieve knowledge advancement about how students' and teachers' learning may benefit each other. In this chapter research on both student learning and teacher learning will be discussed, that departed from a common learning model. Implications for future research that studies student and teacher learning in a more interconnected way will be derived.

S. Krolak-Schwerdt et al., (Eds.), Teachers' Professional Development, 79–95.

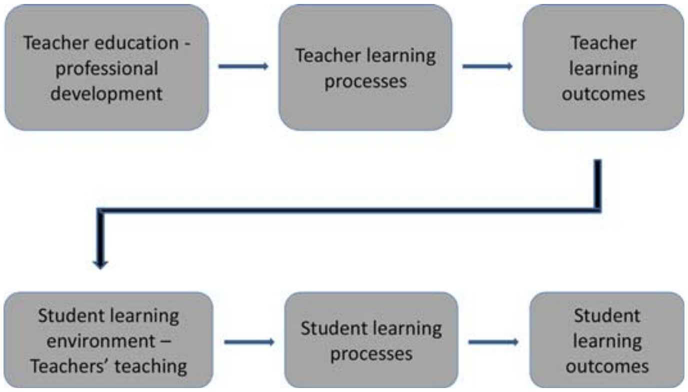
Teacher education - professional development
Teacher learning processes
Teacher learning outcomes
Student learning environment – Teachers' teaching
Student learning processes
Student learning outcomes

Figure 1. Chain of evidence from teacher education and professional development to student learning outcomes.

STUDENT LEARNING

In recent theories of student learning often four domains or components are discerned (see e.g. Entwistle, & McCune, 2004; Lonka, Olkinuora, & Mäkinen, 2004; Pintrich, 2004; Richardson, 2000; Vermunt, & Vermetten, 2004): cognitive processing strategies, metacognitive regulation strategies, conceptions of learning, and learning orientations (see Figure 2).

Cognitive processing strategies are those learning strategies that students use to process the subject matter. They directly lead to learning outcomes in terms of knowledge, understanding, skills, etc. Metacognitive regulation strategies are those learning strategies that students use to regulate and steer their learning processes and lead therefore indirectly to learning outcomes. Conceptions of learning are the beliefs and views people have about learning and associated phenomena: how different learning tasks can be tackled, who is responsible for what in learning, what good teaching looks like, etc. Learning orientations refer to the whole domain of personal goals, motives, expectations, attitudes, worries and doubts students have with regard to learning and studying (Gibbs, Morgan, & Taylor, 1984). Vermunt (1996, 1998) uses the term "learning style" as an encompassing concept in which the cognitive processing of subject matter, the metacognitive regulation of learning, conceptions of learning and learning orientations are united. Later on, because the term "learning style" is often associated with invariant personality characteristics, he and his colleagues changed to using the more neutral term "learning pattern" to denote this united phenomenon (Vermunt, 2005; Vermunt, & Vermetten, 2004). In a series of studies with university students he consistently found four such patterns: undirected, reproduction directed, meaning directed and application directed learning (see also, e.g. Lindblom-Ylänne, 2003; Meyer, 2000).

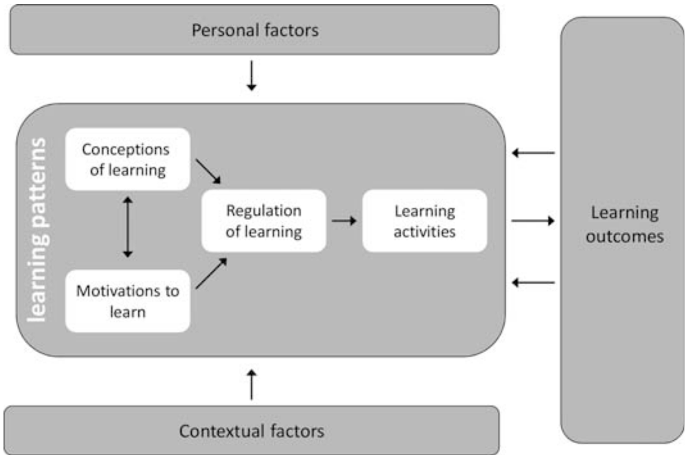

Figure 2. A model of student learning.

First, an *undirected* learning pattern was found, in which students hardly came to processing the subject matter, mainly because they had trouble with selecting what is more and less important within the huge amounts of study materials, showing lack of regulation in their studying, attaching much value to being stimulated by others (fellow-students, teachers, study counselors) in their learning, and having an ambivalent learning orientation showing a lot of doubts about their study choice, own capacities, and the like. Second, a *reproduction directed* way of learning was identified, in which students often used a stepwise processing strategy (combining learning activities like memorizing, rehearsing, detailed analyzing the subject matter), let their learning be regulated by external sources such as teachers and study materials, viewed learning mainly as the intake of knowledge from knowledgeable sources (such as books, teachers), and were certificate and self-test directed in their learning orientation. The third pattern which emerged was a *meaning directed* way of learning, typified by the use of a deep processing strategy (relating, structuring, critical processing the study materials), self-regulation in learning (planning, monitoring, evaluating, reflecting, reading "around" the prescribed materials), a learning conception in which learning was seen as construction of knowledge and one's own responsibility for learning was stressed, as well as personal interest in the subject matter as a learning orientation. And fourth, an *application directed* learning pattern was identified, in which students used a concrete processing strategy (trying to concretize the subject matter, think of possible applications), involved both self and external regulation strategies, attached much importance to learning to use the knowledge they acquired, and were vocation oriented in their learning orientation. Overall, meaning directed learning is mainly focused on relations *within* the subject matter of the studies, application directed learning is focused most on relations *between* the subject matter and the world around.

These four distinctive learning patterns showed up in a number of studies with different student populations. Whether one or the other way of learning is regarded as ‘better’ than another is a matter of perspective. In discussions with teachers and educational developers, meaning and application directed learning are, in general, viewed as more appropriate for studies in higher education than is undirected learning. Sometimes a distinction is made between university and higher vocational studies, in the sense that meaning directed learning is viewed as most appropriate for university studies and application directed learning as most appropriate for higher vocational studies. People often disagree on the value of reproduction directed learning. Some see this as an important route to basic factual knowledge, others argue that this basic factual knowledge can as well, or even be better acquired through meaning or application directed learning (Vermunt, 2007).

CONTEMPORARY EDUCATIONAL INNOVATIONS FOSTERING STUDENT LEARNING

When meaning and application directed learning are ways of learning that are most appropriate or valued, the question arises how we can foster these ways of student learning in our teaching? Most innovative teaching methods that are used nowadays share some common characteristics in the kind of learning they try to promote in students. Overall, these teaching-learning methods aim to foster active, meaning directed, application directed, self-regulated, and cooperative student learning. Main teaching-learning methods used on a large scale nowadays to achieve this aim include assignment based teaching, problem based learning, project-centered learning, competency-based teaching, and dual learning (Vermunt, 2007).

In *assignment-based teaching*, guided self-study is the main learning concept. Compared to traditional teaching there are less lectures, more assignments for self-study and more small tutorial groups. Students conduct their self-study guided by precise instructions in the assignments. In tutorials students’ results of their assignments are discussed and their learning is adjusted by the teaching team. In this way, students actively and independently process the study materials in which they are intensively supervised by the course team (e.g. University of Limburg, 2002). The regulation of students’ learning processes is mainly in the hands of the teachers: they largely decide about the subject matter, learning objectives, criteria for learning outcomes, assessment and feedback. In choosing the learning activities and study resources, students have more freedom and responsibility.

In *problem-based learning (PBL)*, students work in small groups of about ten persons (the tutorial group) trying to understand, explain, and solve problems. The starting point for the learning process is a problem: a short description of a phenomenon about which students should acquire knowledge. In the tutorial group students analyze the problem and formulate learning objectives. After a period of individual study the students meet again and report what they have learned about the problem. There, also unclear matters are clarified and the acquired knowledge is discussed, critically evaluated, and integrated. During their work in the tutorial group the students are guided by a tutor, whose main task is to facilitate the learning and group processes. At the end of a block period, that typically lasts between five to eight weeks, the block test is administered, after which a new block period starts with another theme (see e.g. Dochy, Segers, Van den Bossche, & Gijbels, 2003).

The starting point for *project-centered learning* is a project assignment. This concerns authentic, real-life assignments that are often directly derived from professional practice. Students work in small groups (mostly 4-5 students) independently at the project assignment. Often once a week a meeting takes place under the guidance of a teacher, in which the progress is discussed, difficulties are solved and the next project phase is previewed. The project results in a group product, for example a design, an advice, a plan, a proposition, and the like. At the end of a project block the products or other outcomes are often presented to the whole group of students in the presence of the teachers and sometimes also the customers. The product is assessed along criteria that are made in advance.

A competency is an integrated whole of knowledge, insights, skills, and attitudes. In *competency-based teaching*, students take an intake assessment at the start of their studies, meant to give them insight into their individual starting competencies in relation to the competency profile expected of them at the end of their training. Based on a self-evaluation after a couple of months, they make a personal development plan (PDP), in which they indicate what competencies they will seek to acquire during the remainder of their academic training. In their portfolio, students collect evidence of their growth in the various competencies and they reflect on their development as a whole. When students think they have collected enough evidential material in their portfolio, they can have their portfolio evaluated by an assessor (see e.g. Biemans, Nieuwenhuis, Poell, Mulder, & Wesselink, 2004).

In *dual learning*, students combine two types of learning environments: studying at the university with learning from practice (e.g. Korthagen, 2001). For example, in university-based teacher education programs student teachers do a teaching practice at a secondary school for about half of their study time. There they observe lessons from experienced teachers, conduct lessons themselves, research their practice, supervise pupils, consult their mentor teachers, and form part of the school organization as a whole. The other half of their study program consists of the theoretical part of their studies at the university. Other examples can be found in medical education, law school, etc. The crucial question in such types of training programs is how the different kinds of knowledge students acquire, can be brought together into an integrated knowledge base.

All these teaching-learning methods foster active student learning based on problems, cases or assignments, but they increase in the degree of self-regulation they demand from students and in the magnitude and complexity of the problems on which students work. When they are applied progressively throughout a curriculum, students are challenged to develop their capacity to learn ever more.

The roles that teachers are supposed to fulfill in these different teaching-learning methods may differ considerably. In more traditional teaching, teachers shave to be able to explain the subject matter well, to regulate their students' learning and to motivate students to learn. In assignment-based teaching, skills like making assignments, giving feedback, coaching, and activating students to work are of vital importance. In problem-based learning, the teacher has to be able to fulfill roles like tutor, skills trainer and assessor, and block coordinator. Project-centered learning requires a teacher to supervise project groups, to coach the collaboration within groups, and to deal with students' free-riding behavior. In competency-based teaching, teachers fulfill roles like study career counselor, competency assessor, professional growth consultant, etc. Dual programs capitalize on teacher roles like mentor, portfolio supervisor, authentic test constructor, and being able to clarify students' concerns. In all these innovative teaching-learning methods, teachers have to be able to fulfill roles like diagnostician, challenger, model, activator, monitor, evaluator, and reflector of students' learning processes.

For many teachers these are new roles for which they need new skills that they have not yet mastered. When introducing new teaching methods, there is a huge need for teacher learning which is often heavily underestimated. Many times educational innovations have failed because they did not recognize the need for teacher learning or professional development. New models for teaching students may be developed, fostering active, meaning directed, application directed, self-regulated, and cooperative student learning, but if these methods are too difficult for teachers to use, when they are based on a different conception of teaching and learning than teachers hold, when the principles on which they are based are not understood by teachers, or when teachers do not have the skills to realize the way of teaching in practice, they will not be implemented properly in classroom practices.

Educational innovation succeeds or fails with the teachers that shape it. When an educational institution decides to implement changes in approaches to teaching, teachers are supposed to adapt their way of teaching. This assumes that they learn new things themselves, along the lines of the Shulman and Shulman (2004) model of teacher learning: they are expected to develop another perspective on teaching and learning, need to be willing to learn new skills and play different roles, understand the purpose and application of the innovation, develop skills to realize new teaching in practice, reflect on their experiments with the new teaching methods in order to learn from their experiences, and form part of a community of teachers who all share in learning new things.

SEPARATED AND INTEGRATED TEACHER KNOWLEDGE BASE

Whether being students in teacher education or in-service teachers preparing for educational innovation, teachers have three important sources to learn from: (1) The "theory" of the teacher education or professional development institute. For example, research has been conducted on how students learn, how teachers may best explain certain topics to students, how certain teaching-learning methods may best be used, why students have trouble with certain topics, how the student brain works, etc. Moreover, pedagogical models and learning theories have been developed. Well-educated teachers should know about this scientific or theoretical knowledge. (2) Teachers must build up practical experience with teaching in general or new teaching-learning methods in particular and, by reflecting on those experiences, build up experiential knowledge: about how they may best deal with certain classes or certain types of students, what their own strong and weak sides are, why they should teach differently in school A compared to school B, etc. (3) And, last but not least, pre- and in-service teachers may learn a lot from the rich practical knowledge of experienced teacher educators, mentors, and other teachers in the school (see Figure 3).

Teachers as learners should not only *acquire* knowledge from these three different sources, they also should compare, contrast, and integrate these types of knowledge into a coherent whole. Research has shown that when people do not do this, they develop three more or less independent knowledge bases, in which especially one's own experiential knowledge regulates their actions in classroom practice (Veenman, 1984). These actions are then primarily regulated by knowledge gained from one's own practical experiences, secondly by the practical knowledge gained from experienced colleagues at the school, and thirdly by the "collective" theoretical knowledge gained from the educational institute (compare the thickness of the arrows in Figure 3). This does not seem a desired situation. The central objective in the learning, education and professional development of teachers is thus to bring about an *integrated knowledge base*, in which the three different kinds of knowledge are brought together into one knowledge base and in which this integrated knowledge base, or "theory of practice", regulates teachers' classroom practices (see Figure 4).

TEACHER PROFESSIONAL LEARNING

Figure 5 shows a model of teacher learning patterns in context (cf. Vermunt, & Endedijk, 2011). It is based on the well-researched model of student learning patterns shown in Figure 2. The learning activities that teachers use to learn about their profession are in the center of the model. These learning activities are controlled by regulation activities, which in turn are influenced by teachers' knowledge and beliefs about their own learning (their metacognitive knowledge and beliefs, learning conceptions, etc.) and their learning motivation. Together these four components, and especially the interrelations among these components, constitute a learning pattern. The learning activities that teachers undertake influence the learning outcomes they attain and these learning outcomes are input for new learning processes. Learning patterns are embedded within contextual and person-bound factors, since learning almost always takes place in a social environment. The domain of student learning has a long established tradition in research on patterns in learning (see e.g. Vermunt, & Vermetten, 2004). However, research on teacher learning patterns has only just begun (Vermunt, 2011).

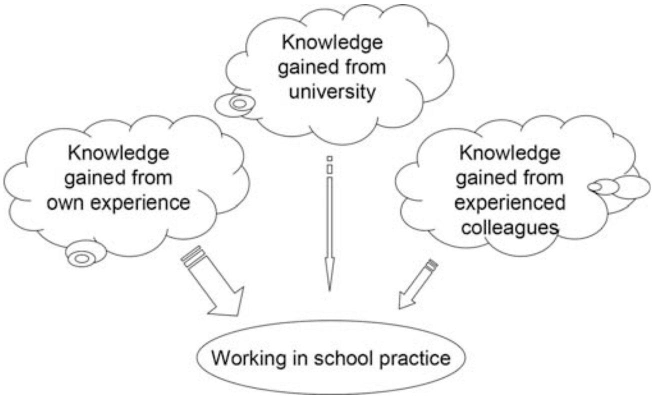
Knowledge gained from university
Knowledge gained from own experience
Knowledge gained from experienced colleagues
Working in school practice

Figure 3. Regulation of working in educational practice by experiential knowledge, theoretical knowledge, and practical knowledge in case of separated knowledge bases.

Mansvelder-Longayroux, Verloop, Beijaard, and Vermunt (2007) studied both student teachers' *learning activities* and self-regulation of learning. They analyzed portfolio segments on the presence of six types of learning activities: remembering, evaluating, analyzing, critical processing, diagnosing, and reflecting. The results showed that 93% of the 1,778 learning activities identified in 39 portfolios referred to remembering and evaluating. "Remembering" in this case meant that an event was described that had occurred in the past (for example something that had happened during a lesson). "Evaluating" meant that a value judgment was attached to the described event (e.g. that went well, wrong, badly, etc.). Only 7% of the learning activities identified in the portfolios referred to a deep approach and self-regulation in learning: analyses, diagnoses, critical processing, or reflecting of or on those events.

Learning of experienced teachers is especially needed when teachers are confronted with innovations in educational practices (James, & McCormick, 2009). In a recent large-scale research project, the central question was how experienced teachers learn from their professional practice in the context of educational reform (Bakkenes, Vermunt, & Wubbels, 2010). During one year the learning experiences of 94 teachers were tracked. The study took place in the context of the introduction of active and self-regulated learning in the classroom, implying a fundamental change in the pedagogical role of the teacher. The teachers were interviewed, observed in the classroom, questionnaires were administered about their beliefs and motives, and six times during the research year they sent in a digital log by email about a learning experience they had had in the previous six weeks. In their logs, they were asked to describe what they had learned, what they wanted to attain, what they had thought and felt, what was the cause of the experience, how they had learned it, what the link was with active and self-regulated learning, and which other people were involved, if any. Consequently, every teacher described about six learning experiences and in total about 500 learning experiences were collected.

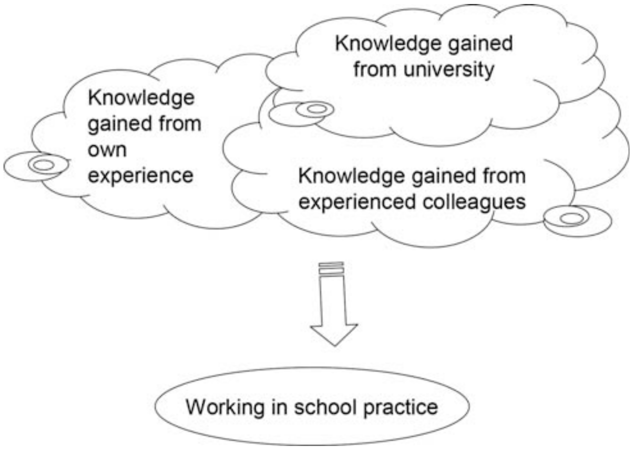
Knowledge gained from university
Knowledge gained from own experience
Knowledge gained from experienced colleagues
Working in school practice

Figure 4. Regulation of working in educational practice by experiential knowledge, theoretical knowledge, and practical knowledge in case of an integrated knowledge base

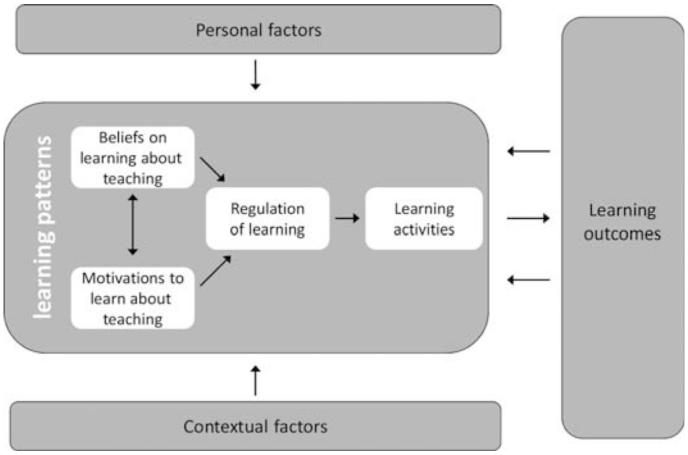

Figure 5. A model of teacher learning.

Bakkenes and colleagues (2010) content-analyzed the learning experiences of the teachers in terms of learning activities and learning outcomes. The former comprised six main categories, namely experimenting (purposefully trying out something new in practice and some form of reflection about it; 32% of all 735 reported learning activities), considering own practice (reflecting on one's own teaching practice and/or students' learning or functioning; 34%), getting ideas from others (taking notice of the views or practices of others and evaluating them; 15%), experiencing friction (noticing a discrepancy between what one expects or wants and what actually happens in class; 15%), struggling not to revert to old ways (trying to change one's teaching but falling back into old routines; 5%), and avoiding learning (engaging in activities that allow one to avoid learning about the new teaching methods; 1%). The first two categories of learning activities were thus reported most frequently.

Regulation processes are assumed to steer the learning activities that teachers use to learn (Butler, Novak Lauscher, Jarvis-Selinger, & Beckingham, 2004; Randi, 2004). Endedijk, Vermunt, Verloop, and Brekelmans (2012) studied the way in which student teachers regulate their learning in a dual teacher education program. These teacher education programs are supposed to greatly appeal on students' self-regulated learning. Until recently, however, self-regulated learning has mainly been studied with students working on scholastic tasks. Endedijk and colleagues (2012) developed an instrument to be able to measure student teachers' self-regulated learning: the "learning report". A content analysis of student teachers' learning reports resulted in the identification and description of a broad variety of regulation activities that student teachers used. The relations among these learning activities could be described with an underlying structure of two dimensions: passive versus active regulation and prospective versus retrospective regulation. Active regulation dominated student teachers' learning in the practice schools, passive regulation dominated their learning at the university. The dimension prospective-retrospective meant that learning experiences that had been started as reactive, spontaneously, and non-planned, could anyway encompass deliberate active regulation activities, but more in a retrospective way. The results also showed that planning and goal setting were no necessary conditions for active regulation of learning in a dual teacher education program.

Van Eekelen, Boshuizen, and Vermunt (2005) studied experienced teacher learning in higher education. The results showed that most (2/3) of the learning experiences that the teachers reported, had not been planned beforehand. The learning activities that were least reported were learning by reading and learning by reflecting, most reported were learning from others and learning by doing. Concretely, this often meant that in the teachers' room ideas were exchanged, that were tried out in class subsequently.

Teachers' knowledge and beliefs about learning and teaching refer to knowledge and beliefs about own learning (conceptions of learning), good teaching, student learning, etc. Many studies show that teachers' knowledge and beliefs exert a strong influence on their classroom practices. On the other hand, relations between cognitions and behavior are not always as straightforward as one would think (Meirink, Meijer, Verloop, & Bergen, 2009). Tigchelaar, Brouwer, and Vermunt (2010) found that second-career teachers often have developed strong beliefs about learning and teaching, rooted in earlier experiences that exert a strong influence on both their own teaching and learning.

A considerable number of studies has been conducted on teachers' beliefs about (good) teaching. For example, Postareff and Lindblom-Ylänne (2008) interviewed teachers from various academic disciplines about their views on good teaching.

They found two broad categories: a view focusing on student learning and a view focusing on the subject matter. Many other authors reported a comparable distinction between two fundamental views of teachers about teaching, sometimes using a slightly different terminology, such as a teacher-oriented versus a student-oriented view on teaching (e.g. Entwistle, 2009; Samuelowicz, & Bain, 2001; Trigwell, & Prosser, 2004). Studies on teachers' views about their own learning are rare, an exception being the study of Boulton-Lewis, Wilss, and Mutch (1996). They studied the conceptions and knowledge of 40 experienced teachers, who were doing an in-service course as adult students, about their own learning. A remarkable finding was that these teachers hardly mentioned any factors that can be found in the literature as characteristic for adult students. The authors suggest that even teachers should learn more about their own learning to be able to learn in a self-directed way.

In the research literature a considerable number of studies can be found about teachers' *motivation for teaching* and about how teachers can motivate students to learn (e.g. Kocabas, 2009; Woolfolk-Hoy, 2008). Much less research has been done on teachers own motivation to learn. Watts and Richardson (2008) studied student teachers' motivation to learn to teach and found three types: highly committed persisters who planned to be a teacher for their whole life; highly committed switchers who had other career plans than staying a teacher for their whole life; and lowly committed desisters who had become dissatisfied with the choice for a teaching career and were not intending to follow-up this initial choice. Van Eekelen, Vermunt, and Boshuizen (2006) studied the will to learn of experienced teachers who were confronted with educational reforms. They identified three forms of teachers' will to learn: (1) teachers who did not see why they should learn something new; (2) teachers who wondered how they should learn something new; and (3) teachers who were eager to learn new things. These last teachers were open towards possible learning situations as work, alert to what happened in the classroom and the needs and behavior of individual students, critical on their own functioning, they tried to improve their practice, and they could indicate what they had learned from their work. The second group showed these characteristics to a lesser degree, the first group hardly showed any of these characteristics.

In research on student learning, *learning outcomes* are often simply operationalized as scores on a knowledge test. The conceptualization of learning outcomes in teacher learning should be more advanced. Bakkenes and colleagues (2010) content analyzed the learning outcomes that experienced teachers reported in the context of educational innovations and they could identify the following four types: (1) changes in knowledge and beliefs (being more aware of something, confirmation of existing ideas, developed new ideas); (2) intentions for practice (try out new practices, continue new practices, continue current practices); (3) changes in teaching practices in a more permanent way (new practices in line with the innovation, back to old practices); and (4) changes in emotions (positive emotions such as pride, satisfaction, hope; negative emotions as irritation, anger, shock, fear; emotions of surprise). The teachers in this study mostly reported changes in knowledge and beliefs (50.0% of all

1287 learning outcomes they reported), followed by changes in emotions (35.0%), intentions for practice (13.5%) and changes in classroom practices (1.4%). Learning activities turned out to be associated with all measures of learning outcomes. For example, the learning activity "experimenting" turned out to be associated with the learning outcomes "intention to continue new practices", "confirmation of beliefs", "positive emotions", and "surprises".

The most direct *contextual factor* assumed to influence teacher learning is the learning environment. In teacher education this learning environment encompasses, for example, the educational program attended at the institute, but also the practice school. In an in-service professional development program the learning environment encompasses the type of intervention that is used, the social environment, but also the wider school climate (openness to innovation, learning orientedness, etc.), and other contextual variables (Clarke, & Hollingsworth, 2002).

Personal factors are, for example, personality characteristics, teaching and learning experience, professional identity, well-being, age, gender, and school subject taught. In their study among teacher education students, Oosterheert, Vermunt, and Veenstra (2002) found that patterns in learning to teach were associated with various personal and contextual variables. Personality characteristics that turned out to be associated with meaning oriented learning were self-confidence, extraversion, emotional stability, and tolerance for uncertainty. Learning patterns proved not to be associated with personal factors as age, gender, subject taught, number of teaching hours a week, and educational experience outside of formal education. With regard to contextual factors Oosterheert and colleagues (2002) found that meaning directed students perceived their learning environment as more constructive than students with other learning patterns.

The degree to which teachers' adaptation to an educational innovation can be explained by personal and contextual factors was studied by Vermunt, Bakkenes, Wubbels, and Brekelmans (2008). Personal factors in this study referred to, among other things, learning motivation, learning conceptions, professional identity, and personality characteristics. Contextual factors included type of learning environment, dominant beliefs in the school, and organizational climate. Teaching practices were measured in terms of student perceptions. The results showed that person-bound factors were more important than contextual factors in explaining teachers' beliefs and practices. For explaining learning activities and learning outcomes, however, contextual factors turned out to be more important than personal factors. The most important contextual factor was the type of learning environment. Organized learning environments (peer coaching, collaboration in teams) turned out to elicit qualitatively better learning activities and learning outcomes than informal workplace learning (Bakkenes et al., 2010).

Oosterheert and Vermunt (2001) interviewed student teachers extensively about their learning activities, the way in which they combined theory and practice, the way in which they regulated their learning, their beliefs about learning, and their worries, doubts, and emotions in learning. They could identify five *patterns in learning* to teach: two variants of a meaning oriented pattern, two variants of a reproduction oriented pattern, and a survival oriented pattern. Survival oriented student teachers thought that a lot of teaching automatically led to learning the profession. They learned a lot by doing without having clear learning goals in mind. Reproduction oriented student teachers tried to learn the profession by trying out teaching activities in practice. They tried to remember what worked, to forget what did not work, were strongly focused on their performance in the classroom, and tried to improve their teaching through collecting ready-made practical tips. Meaning oriented student teachers thought that learning to teach mainly came down to developing their frame of reference, a better understanding of teaching and learning. Some of them needed a lot of external support, others themselves actively combined information from different sources to build up an integrated knowledge base. Donche and Van Petegem (2009), in their study among Flemish teacher education students, found comparable patterns in learning to teach as the ones found by Oostheert and Vermunt (2001) among Dutch student teachers.

Some recent studies have revealed the beginnings of patterns in the learning of experienced teachers (Bakkenes et al., 2010; Hoekstra, Brekelmans, Beijaard, & Korthagen, 2009; Vermunt, 2011). For example, Bakkenes and colleagues (2010) found that most teachers' learning was sharply focused upon improving their *immediate performance* in class: they wanted to be able to use what they had learned quickly in their teaching. Another group of teachers was (also) *meaning directed* in their learning: this group wanted to know why things in class work as they work, looked for reasons behind it, worked at extension of their theory of practice, often involved knowledge and "theory" from outside, and often worked for a longer period of time on a certain learning or development theme. A third group, substantial in size, could be characterized as *undirected* in their learning. These teachers struggled with the educational innovation, often did not know well how they could teach in another way, and did not know as well how they could *learn* about teaching in another way (Vermunt, & Endedijk, 2011).

CONCLUSIONS AND DISCUSSION

Comparing and contrasting the results of the research on student learning and teacher learning, it can be concluded that in both student learning and teacher learning *meaning oriented* learning and *undirected* learning play a major role. These dimensions are important aspects of models conceptualizing both student and teacher learning, although the way these phenomena are manifested in concrete appearance differ for both populations. For example, meaning orientation in student learning manifests itself as the search for understanding relations between different parts of the study texts and between different courses and trying to structure one's ideas and thoughts into a personal understanding of the literature. However, teachers do not learn predominantly by studying books. Meaning orientation in teacher learning is manifested more by a search for understanding relations between different classroom experiences they have had, between new ideas and own experiences, and by trying to structure one's reflections into a personal theory of practice. *Application directed* learning, identified as an important dimension in student learning especially among older and advanced students, resembles the way of learning among teachers directed towards *immediate performance* improvement in the classroom. *Reproduction directed* learning, a major dimension in student learning, did not have an equivalent in teacher learning. It seems that this type of learning is typical for students' learning as long as they are within educational boundaries, but after they leave the school it is no longer considered a useful way of learning.

Education that is aimed at teaching students to learn and think in an increasingly self-regulated way is characterized by a gradual shift in the task division in the learning process from educational "agents" (e.g. teacher, tutor, book, or computer) to students (Vermunt, & Verschaffel, 2000). Initially, explicit external regulation is offered to students. Subsequently, that support is gradually withdrawn. At the same time, students are taught how to exert control over their learning processes themselves. Learning to learn and think independently means a gradual transfer of learning activities from the teachers to the students, a gradual shift from external to internal regulation of learning. The method of teaching typically changes to allow increasing self-regulation in students' learning. As a result, students are continuously challenged to try a next step in their self-regulated learning and thinking. The ultimate goal is to help students become life-long learners by the time they graduate, willing and able to keep on developing in their professional area after the termination of their educational career and to never stop learning.

In practice, however, teaching-learning methods do not change much over time to encourage greater student self-regulation. If we consider it important that students learn to learn and think in an ever more self-regulated way, the teaching-learning methods need to change to provide a progressive increase in self-regulation year by year (e.g. Ten Cate, Snell, Mann, & Vermunt, 2004). This would then lead to a curriculum that is typified by gradual, systematically decreasing external regulation from teachers and an increasing self-regulation by students (transfer of regulation). That might lead to a pattern of teaching-learning methods that succeed each other: for example, assignment-based teaching could be the starting point followed by PBL, project-centered learning, competency-based teaching, and dual learning.

Just as first-year students find it difficult to adjust to the self-regulation required in innovatively taught courses, so teachers find it equally difficult to adopt the very different roles that educational developers are expecting of them in innovative teaching methods. The transitions are not simply of learning new skills, but fundamentally changing a mindset that previously involved regulating the study program and controlling student activities working independently, to one which accepts an increasing level of student autonomy and collaborative learning. To call this change "demanding" is to understate what is being expected; any such change will take many years to accomplish, but such changes are essential if we are to help students to become self-regulated and self-motivated learners by the time they leave school.

Findings as reported in this chapter have significant implications for educational practice. They contributed to our understanding of how teachers learn and this knowledge can be crucial for designing powerful environments to foster teacher learning. Until now, attempts to foster teacher learning or professional development have been characterized by a high degree of "beliefs". The various institutes and agencies responsible for teacher professional development all strongly believe in their own approaches, while at the same time these approaches are very diverse. There is only limited scientific evidence to support claims for the effectiveness of any of these different approaches (Grossman, 2005). As in the field of student learning, we are convinced that any theory or model of fostering teacher learning and professional development should be based on research evidence of how teachers learn (cf. Beijaard, Korthagen, & Verloop, 2007). The studies discussed here aimed to contribute to the scientific knowledge base of teacher learning.

Future research should be directed at further scrutiny of the patterns of teacher learning found in the studies discussed above. The components of the model of teacher learning and their interrelations should be studied with different teacher populations and different stages of expertise. Moreover, there is an urgent need to develop and research diagnostic instruments for measuring the various elements of teacher learning. Last but not least, we also need studies directed at testing and further developing pedagogical approaches to foster high quality teacher learning. Developing intervention models, based on scientific evidence that can support and foster teacher learning in the context of educational innovation, and studying the power and effects of these models in bringing about teacher learning, are in our view important tasks for educational research in this field for the years to come.

REFERENCES

Bakkenes, I., Vermunt, J. D., & Wubbels, T. (2010). Teacher learning in the context of educational innovation: Learning activities and learning outcomes of experienced teachers. *Learning and Instruction, 20,* 533–548.

Beijaard, D., Korthagen, F., & Verloop, N. (2007). Understanding how teachers learn as a prerequisite for promoting teacher learning. *Teachers and Teaching: Theory and Practice, 13,* 105–108.

Biemans, H., Nieuwenhuis, L., Poell, R., Mulder, M., & Wesselink, R. (2004). Competence-based VET in The Netherlands: Background and pitfalls. *Journal of Vocational Education and Training, 56*, 523–538.

Boulton-Lewis, G. M., Wilss, L., & Mutch, S. (1996). Teachers as adult learners: Their knowledge of their own learning and implications for teaching. *Higher Education, 32*, 89–106.

Butler, D. L., Novak Lauscher, H., Jarvis-Selinger, S., & Beckingham, B. (2004). Collaboration and self-regulation in teachers' professional development. *Teaching and Teacher Education, 20*, 435–455.

Clarke, D., & Hollingsworth, H. (2002). Elaborating a model of teacher professional growth. *Teaching and Teacher Education, 18,* 947–967.

Dochy, F., Segers, M., Van den Bossche, P., & Gijbels, D. (2003). Effects of problem-based learning: A meta-analysis. *Learning and Instruction, 13*, 533–568.

Donche, V., & Van Petegem, P. (2009). The development of learning patterns of student teachers: A cross-sectional and longitudinal study. *Higher Education, 57*, 463–475.

Endedijk, M., Vermunt, J. D., Verloop, N., & Brekelmans, M. (2012). The nature of student teachers' regulation of learning in teacher education. *British Journal of Educational Psychology, 82,* 469–491.

Entwistle, N. J. (2009). *Teaching for understanding at university.* Basingstoke, UK: Palgrave Macmillan.

Entwistle, N. J., & McCune, V. (2004). The conceptual bases of study strategy inventories. *Educational Psychology Review, 16,* 325–345.

Gibbs, G., Morgan, A., & Taylor, E. (1984). The world of the learner. In F. Marton, D. Hounsell, & N. J. Entwistle (Eds.), *The experience of learning* (pp. 165–188). Edinburgh, UK: Scottish Academic Press.

Grossman, P. (2005). Research on pedagogical approaches in teacher education. In M. Cochran-Smith, & K. M. Zeichner (Eds.), *Studying teacher education – The report of the AERA panel on research and teacher education* (pp. 425–476). Mahwah, NJ: Erlbaum.

Hoekstra, A., Brekelmans, M., Beijaard, D., & Korthagen, F. (2009). Experienced teachers' informal learning: Learning activities and changes in behavior and cognition. *Teaching and Teacher Education, 25*, 663–673.

James, M., & McCormick, R. (2009). Teachers learning how to learn. *Teaching and Teacher Education, 25,* 973–982.

Kocabas, I. (2009). The effects of sources of motivation on teachers' motivation levels. *Education, 129,* 724–733.

Korthagen, F. (2001). *Linking practice and theory. The pedagogy of realistic teacher education.* Mahwah, NJ: Erlbaum.

Lindblom-Ylänne, S. (2003). Broadening an understanding of the phenomenon of dissonance. *Studies in Higher Education, 28,* 63–78.

Lonka, K., Olkinuora, E., & Mäkinen, J. (2004). Aspects and prospects of measuring studying and learning in higher education. *Educational Psychology Review, 16*, 301–323.

Mansvelder-Longayroux, D., Verloop, N., Beijaard, D., & Vermunt, J. D. (2007). Functions of the learning portfolio in student teachers' learning process. *Teachers College Record, 109*, 126–159.

Meirink, J. A., Meijer, P. C., Verloop, N., & Bergen, T. C. M. (2009). Understanding teacher learning in secondary education: The relations of teacher activities to changed beliefs about teaching and learning. *Teaching and Teacher Education, 25,* 89–100.

Meyer, J. H. F. (2000). An empirical approach to the modelling of dissonant study orchestrations in higher education. *European Journal of Psychology of Education, 15,* 5–18.

Oosterheert, I. E., & Vermunt, J. D. (2001). Individual differences in learning to teach – Relating cognition, regulation and affect. *Learning and Instruction, 11*, 133–156.

Oosterheert, I. E., Vermunt, J. D., & Veenstra, D. R. (2002). Manieren van leren onderwijzen en relaties met persoonsgebonden en contextuele variabelen [Orientations towards learning to teach and relations to personal and contextual variables]. *Pedagogische Studiën, 79*, 251–268.

Pintrich, P. (2004). A conceptual framework for assessing motivation and self-regulated learning in college students. *Educational Psychology Review, 16,* 385–408.

Postareff, L., & Lindblom-Ylänne, S. (2008). Variation in teachers' descriptions of teaching: Broadening the understanding of teaching in higher education. *Learning and Instruction, 18*, 109–120.

Randi, J. (2004). Teachers as self-regulated learners. *Teachers College Record, 106*, 1825–1853.

Richardson, J. T. E. (2000). *Researching student learning.* Buckingham, UK: SRHE and Open University Press.

Samuelowicz, K., & Bain, J. D. (2001). Revisiting academics' beliefs about teaching and learning. *Higher Education, 41,* 299–325.

Shulman, L. S., & Shulman, J. H. (2004). How and what teachers learn: a shifting perspective. *Journal of Curriculum Studies, 36*, 257–271.

Ten Cate, O., Snell, L., Mann, K., & Vermunt, J. (2004). Orienting teaching towards the learning process. *Academic Medicine, 79*, 219–228.

Tigchelaar, A., Brouwer, N., & Vermunt, J. D. (2010). Tailor made: Towards a pedagogy for educating second-career teachers. *Educational Research Review, 5,* 164–183.

Trigwell, K., & Prosser, M. (2004). Development and use of the Approaches to Teaching Inventory. *Educational Psychology Review, 16,* 409–424.

University of Limburg (2002). *Onderwijsontwikkelingsplan LUC/tUL. Bijlage 2: Actieplannen van faculteiten en schools.* [Educational development plan University of Limburg. Appendix 2: Action plans of faculties and schools]. Diepenbeek, Belgium: University of Limburg.

Van Eekelen, I. M., Boshuizen, H. P. A., & Vermunt, J. D. (2005). Self-regulation in higher education teacher learning. *Higher Education, 50*, 447–471.

Van Eekelen, I. M., Vermunt, J. D., & Boshuizen, H. P. A. (2006). Exploring teachers' will to learn. *Teaching and Teacher Education, 22*, 408–423.

Veenman, S. (1984). Perceived problems of beginning teachers. *Review of Educational Research, 54,* 143–178.

Vermunt, J. D. (1996). Metacognitive, cognitive and affective aspects of learning styles and strategies: A phenomenographic analysis. *Higher Education, 31,* 25–50.

Vermunt, J. D. (1998). The regulation of constructive learning processes*. British Journal of Educational Psychology, 68,* 149–171.

Vermunt, J. D. (2005). Relations between student learning patterns and personal and contextual factors and academic performance. *Higher Education, 49,* 205–234.

Vermunt, J. D. (2007). The power of teaching-learning environments to influence student learning. *British Journal of Educational Psychology Monograph Series II, 4*, 73–90.

Vermunt, J. D. (2011). Patterns in student learning and teacher learning: Similarities and differences. In S. Rayner, & E. Cools (Eds.), *Style differences in cognition, learning and management: Theory, research and practice* (pp. 173–187). New York: Routledge.

Vermunt, J. D., & Endedijk, M. D. (2011). Patterns in teacher learning in different phases of the professional career. *Learning and Individual Differences, 21*(3), 294–302.

Vermunt, J. D., & Vermetten, Y. J. (2004). Patterns in student learning: Relationships between learning strategies, conceptions of learning, and learning orientations. *Educational Psychology Review, 16*, 359–384.

Vermunt, J. D., & Verschaffel, L. (2000). Process-oriented teaching. In R. J. Simons, J. van der Linden, & T. Duffy (Eds.), *New Learning* (pp. 209–225). Dordrecht, The Netherlands: Kluwer Academic Publishers.

Vermunt, J. D., Bakkenes, I., Wubbels, T., & Brekelmans, M. (2008, August 21–23). *Personal and contextual factors and secondary school teachers' adaptation of innovation.* Paper presented at the 11th International Conference on Motivation, Turku, Finland.

Watts, H. M. G., & Richardson, P. W. (2008). Motivations, perceptions, and aspirations concerning teaching as a career for different types of beginning teachers. *Learning and Instruction, 18,* 408–428.

Woolfolk-Hoy, A. (2008). What motivates teachers? Important work on a complex question. *Learning and Instruction 18,* 492–498.

AFFILIATIONS

Jan D. Vermunt
University of Cambridge

DIRK RICHTER, MAREIKE KUNTER, UTA KLUSMANN,
OLIVER LÜDTKE & JÜRGEN BAUMERT

PROFESSIONAL DEVELOPMENT ACROSS THE TEACHING CAREER: TEACHERS' UPTAKE OF FORMAL AND INFORMAL LEARNING OPPORTUNITIES[*]

INTRODUCTION

Over the last decade, the debate on school quality (U.S. Congress, 2001) has focused increased attention on teachers' professional development. According to the National Commission on Teaching and America's Future's (2003), "strong professional development opportunities must be embedded in the very fabric of public education" (p. 129). Modern views of professional development characterise professional learning not as a short-term intervention, but as a long-term process extending from teacher education at university to in-service training at the workplace (Ball & Cohen, 1999; Feiman-Nemser, 2001; Putnam & Borko, 2000). Although recent educational reforms support the concept of lifelong learning (U.S. Congress, 2001), little is known about the learning opportunities used by teachers or about how patterns of uptake change across the teaching career (Corcoran, 2007). Empirical studies investigating teachers' participation in professional development have identified age-related differences (Desimone, Smith, & Ueno, 2006; Mesler & Spillane, 2009; Mok & Kwon, 1999), but without putting these in the context of teacher career development. The literature suggests that the teaching career can be divided into consecutive stages with different implications for participation in professional development (Huberman, 1989; Sikes, Measor, & Woods, 1985). Although the empirical basis is rather weak, findings indicate that beginning teachers tend to use observations and informal discussions with colleagues to improve their practice, whereas more experienced teachers are more inclined to use formal meetings for their professional learning (Grangeat & Gray, 2007). In other words, teachers seem to use different learning opportunities across the career cycle. However, empirical studies have not systematically explored how teachers of different ages make use of professional development opportunities.

This study investigates teachers' uptake of learning opportunities across the career. First, we examine changes in the uptake of formal and informal learning opportunities. Second, we focus on the content of the formal learning opportunities and explore whether their uptake differs with respect to content. Finally, we seek

S. Krolak-Schwerdt et al., (Eds.), Teachers' Professional Development, 97–121.

to explain career-related changes in the uptake of formal and informal learning opportunities by reference to teachers' work engagement and additional professional responsibilities within the school.

PROFESSIONAL DEVELOPMENT AS UPTAKE OF DIFFERENT KINDS OF LEARNING OPPORTUNITIES

We define professional development as uptake of *formal* and *informal learning opportunities* that deepen and extend teachers' professional competence, including knowledge, beliefs, motivation and self-regulatory skills (Baumert & Kunter, 2006; Kunter et al., 2007). This definition distinguishes between formal and informal learning opportunities (Desimone, 2009). *Formal learning opportunities* are defined as structured learning environments with a specified curriculum, such as graduate courses or mandated staff development (Feiman-Nemser, 2001). They represent a main component of the "training model" (Little, 1993, p. 129), also known as the "traditional view" (Lieberman, 1995, p. 591) on professional development. The training model assumes that teachers update their knowledge and skills by means of workshops and courses. These are generally full- or half-day activities in which experts disseminate information that can be applied in the workplace (Feiman-Nemser, 2001). Many European countries and U.S. states require their teachers to attend such activities on a regular basis. As such, they are still the most widely used form of professional development (Eurydice, 2008; National Association of State Directors of Teacher Education and Certification (NASDTEC), 2004). According to Hill (2007), teachers in most U.S. states need to complete 120 h of professional development per 5-year period. European professional development requirements vary greatly, ranging from 12 to 57 h per year (Eurydice, 2008).

Informal learning opportunities, in contrast, do not follow a specified curriculum and are not restricted to certain environments (Desimone, 2009). They include individual activities such as reading books and classroom observations as well as collaborative activities such as conversations with colleagues and parents, mentoring activities, teacher networks and study groups (Desimone, 2009; Mesler & Spillane, 2009). Participation in these activities is generally not mandatory (Eurydice, 2008; NASDTEC, 2004), but is at teachers' own initiative. As such, teachers are not merely recipients of knowledge. Rather, they organise the learning process and determine their learning goals and strategies independently. Moreover, informal learning opportunities are often embedded in the classroom or school context, which allows teachers to reflect on their practice and to learn from their colleagues (Putnam & Borko, 2000). Of the broad variety of informal learning opportunities, this study focuses on teacher collaboration and the use of professional literature. Teacher collaboration was selected as an indicator of informal learning in the professional community. It is situated in the school and classroom context and is based on interactions with colleagues (Goddard, Goddard, & Tschannen-Moran, 2007; Putnam & Borko, 2000). This kind of learning is a cooperative process in which teachers come together to discuss and share knowledge (Lieberman, 1995; Putnam & Borko, 2000), learning from each other's experiences and gaining new insights into teaching and learning (Putnam & Borko, 2000). In addition to collaborative learning, we were also interested in examining individual learning activities beyond the school context. One example is the use of the professional literature: subject matter literature, journals and teaching materials (Kwakman, 2003; Scribner, 1999). Professional literature can serve as a resource for instructional activities, offer information about educational policy and suggest ways of dealing with work-related demands (Scribner, 1999). In this study, we are interested in teachers' use of professional literature as an informal learning activity that can be engaged in independently, without much planning and at the teacher's own pace.

UPTAKE OF LEARNING OPPORTUNITIES

The largest national study to have examined teachers' uptake of different learning opportunities is the U.S. Schools and Staffing Survey, 1999-2000 (SASS; Choy, Chen, & Bugarin, 2006). The study was based on a representative sample of approximately 54 200 school teachers and provides insights into the professional development activities of the U.S. teaching force (U.S. Department of Education & National Center for Education Statistics, 2002). Teachers in public and private schools were asked about their uptake of various learning opportunities over a 12-month period. Findings showed that 98.3% of teachers participated in some type of learning opportunity, but that most of them used formal activities such as workshops, conferences or training courses (Choy et al., 2006). Fewer teachers participated in informal activities such as regularly scheduled collaboration with colleagues (72.6%), individual or collaborative research (46.5%) or mentoring and peer observation (41.9%). In sum, there is a great deal of variation in teachers' participation in different learning opportunities. These differences can be attributed to differences in both the availability of professional development activities and their uptake by teachers (Cookson, 1986). Previous empirical studies investigating the uptake of professional development activities have focused on the impact of individual teacher characteristics (i.e., motivation and beliefs) and characteristics of the work context (i.e., work load, cf. Lohman, 2000; Mok & Kwon, 1999). However, this focus on inter-individual differences neglects the fact that individuals' professional development behaviour changes over time. Although teachers in many states are required to participate in professional development on a continuous basis, little is known about the use of learning opportunities across the teaching career.

TEACHER CAREER STAGE MODELS AS A FRAMEWORK FOR A LIFESPAN PERSPECTIVE ON PROFESSIONAL DEVELOPMENT

The primary goal of teacher career stage models is to describe the prototypical development of individual teacher characteristics in terms of discrete stages (see overview in Fessler, 1995). The first models were introduced in the 1970s (Gregorc, 1973; Unruh & Turner, 1970), with many others being proposed in the 1980s (e.g., Huberman, 1989; Sikes et al., 1985). Career stage models have been re-visited in recent years, and new models have been proposed (Dall'Alba & Sandberg, 2006; Day, Sammons, Stobart, Kington, & Gu, 2007). The stages identified represent common aspects of individual development (in terms of, e.g., knowledge, skills and goals) as well as teachers' position within the school community and the wider profession. The complexity of the models makes it possible to derive hypotheses about implications for teachers' professional development. We selected Huberman's (1989) career stage model as a theoretical framework for the present study because it is widely accepted in the literature and provides an in-depth description of the teaching career from start to finish. The model has frequently been used to interpret the results of empirical studies from a life- span perspective (Anderson & Olsen, 2006; Brekelmans, Wubbels, & van Tartwijk, 2005; Choi & Tang, 2009). To date, however, no comprehensive validation studies of the model have been performed. Our study does not aim at validating the entire model. Rather, we apply it as a heuristic to derive and examine hypotheses about the uptake of learning opportunities during the teaching career.

The Huberman model characterises development as a set of five consecutive stages (i.e., survival and discovery, stabilisation, experimentation/activism and stocktaking, serenity and conservatism, and disengagement) which are closely connected to individual teaching experience. These stages represent major phases of teachers' development, but they do not necessarily apply to each and every teacher in the same way. In the following, we discuss the individual stages and illustrate them with empirical data from the SASS (1999-2000) report.

The Beginning of the Career: The First and Second Phase

The first 3 years of the teaching profession are a time of "survival and discovery". Beginning teachers experience their initial years of teaching as a struggle for survival (see also Day et al., 2007; Veenman, 1984), typically reporting a sense of exhaustion, feeling overwhelmed, problems with student discipline and continuous trial and error. At the same time, fulfilling the responsibilities of a classroom teacher brings a sense of accomplishment and discovery. The second, "stabilisation" phase occurs around years 4-6. During this time, teachers become more established in their profession (e.g., by obtaining tenure) and more affiliated with the teaching community. They also develop and further refine their instructional skills. But how do teachers learn during this period? Many U.S. states and some European countries (i.e., Austria, England and Germany) provide induction programs for beginning teachers (Corcoran, 2007; Eurydice, 2002). These programs are developed to ease the transition from college to classroom teaching by offering beginning teachers with formal and informal learning opportunities, including mentors, study groups, classroom observations and formal professional development activities (Glazerman et al., 2008). Moreover, beginning teachers learn through independent classroom teaching and informal discussions with other teachers (Grangeat & Gray, 2007).

Empirical data from the SASS (1999-2000) show that beginning teachers (1-3 years of experience) participated more frequently than any other group of public school teachers in mentoring or peer observation (50.7% over the 12-month period surveyed). They also showed high attendance of formal activities such as conferences and workshops (93.3%) and continued to attend university courses in their main teaching subject (31.5%). In other words, many beginning teachers participated in informal activities while continuing their formal training. In terms of the content of the activities pursued, beginning teachers attended more activities targeting classroom management and student discipline than did experienced teachers (more than 3 years of experience). Beginning teachers thus chose to attend activities dealing with topics that are particularly challenging for those new to the profession. The second phase of the model cannot be illustrated with data from the SASS (1999-2000) report, because the report does not specify groups of teachers that fully match the stages of the Huberman model.

The Middle of the Career: The Third and Fourth Phase

The third phase – "experimentation/activism and stock- taking" – covers years 7-18 of the career and has two possible orientations: (1) "experimentation and activism" or (2) "reassessment and self-doubts". Teachers who wish to increase their instructional impact may experiment with new materials and instructional strategies (Huberman, 1989). This activism may carry over into the school community and lead to additional professional responsibilities or promotion (e.g., coordinator, department head). In the absence of experimentation, teachers may experience self-doubts and consider leaving the profession. The fourth phase covers years 19 and 30 of the career and again has two possible orientations: (1) "serenity" or (2) "conservatism". Serene teachers experience a loss of engagement, a decline in career ambitions, but also greater sense of self-acceptance, whereas conservative teachers are sceptical towards educational innovations and critical of educational policy (Peterson, 1964). To what extent do teachers continue learning in this period of the career? Teachers at this stage are already very experienced; however, the Huberman (1989) model predicts that they remain very interested in adding to their knowledge and skills. Data from the SASS (1999-2000) show that public school teachers with 10-19 years of experience were more involved in regularly scheduled collaboration (77.9%), individual collaborative research (48.9%) and observational visits to other schools (36.4%) than were beginning or more experienced teachers (Choy et al., 2006). Moreover, they frequently attended formal workshops and conferences (95.8%). In terms of content, this group of teachers participated most intensively in activities relating to their teaching subject, content and performance standards and teaching methods. These figures support the theoretical proposition that teachers in this phase of their career continue to pursue various professional development activities in order to develop their instructional repertoire.

The End of the Career: The Fifth Phase

The last phase of the teaching career begins with approximately 30 years of teaching experience and is characterised by withdrawal from the profession. Teachers at this stage tend to reduce their commitment and career ambition, instead focusing more on personal goals. This change in career motivation can also be linked to teachers' use of learning opportunities. As the potential future return on professional development activities decreases when retirement is imminent, teachers at this point in their career can be expected to reduce their involvement in these activities. Indeed, empirical data from the SASS (1999-2000) demonstrate that across all learning opportunities investigated public school teachers with more than 20 years of experience exhibited lower participation rates than did mid-career teachers (Choy et al., 2006). In particular, they showed the lowest rates of participation in university courses in their main teaching subject (16.4%) and in university courses leading to re-certification and advanced certification (21.2%). In other words, only a small group of teachers in this phase pursued further qualifications. With respect to content, this group of teachers showed less participation in activities pertaining to their subject, content and performance standards, teaching methods, student discipline or classroom management. However, they showed increased participation in training activities on the use of computers for instruction and student assessment. These results highlight two important findings: In accordance with the model of career development, teachers' engagement in professional development declines at this stage of the career. However, their average involvement in content areas affected by recent technological developments and educational reforms increases (U.S. Congress, 2001).

THE PRESENT STUDY

Huberman's (1989) theoretical framework of teachers' career stages suggests that teachers make use of different types of learning opportunities across their careers. Existing data have provided first insights into differential patterns of formal and informal learning opportunities, but no previous study has explicitly examined the relationship between teachers' age and participation in professional development. Thus, it is the aim of this study to investigate teachers' uptake of different learning opportunities from the beginning to the end of the teaching career. The first part of the present study examines the differential uptake of formal and informal learning opportunities. We focus on in-service training as an example of formal learning opportunities and on teacher collaboration and the use of professional literature as two examples of informal learning opportunities. Based on the career stage model, we hypothesise that teachers pursue in-service training most frequently in the middle of their careers and show less involvement as they approach retirement. Similarly, we predict that teachers collaborate most intensively in the middle of their careers, reducing their involvement thereafter. The use of professional literature across the career has not yet been investigated, but we speculate that teachers read less professional literature towards the end of their careers, as the potential payoffs of this activity also decrease. The second part of the study addresses the previous finding that teachers with different levels of experience select professional development activities in different content areas (Choy et al., 2006). More specifically, teachers tend to choose activities that reflect the demands or professional goals of the career stage they are in (e.g., beginning teachers are more likely to seek further training in classroom management). To further investigate these relationships, we developed a categorisation scheme to classify the content of teachers' formal learning opportunities. Based on these categorisations, we investigate whether content-specific learning opportunities are used differently across the teaching career. The third part of the study seeks to identify individual teacher characteristics that predict career-related change in the uptake of formal and informal learning opportunities. As described above, teachers' motivation for professional advancement is strongest at mid-career and drops off towards retirement. Furthermore, research on professional development has shown that teachers who seek promotion (Mok & Kwon, 1999) or are interested in developing professionally (Kwakman, 2003) use learning opportunities more frequently. Thus, we hypothesise that teachers' age is not the primary factor explaining changes in learning behaviour, but that it serves as a proxy for motivational variables. We therefore predict that teachers' work engagement (i.e., a motivational disposition to progress in the career and willingness to invest resources to achieve this goal) explains age-related differences in the use of learning opportunities. In addition, we predict that teachers' use of learning opportunities is related to their professional responsibilities. Teachers with additional professional responsibilities and duties in the school community (e.g., in school administration) may be more likely to participate in learning opportunities. We therefore examine whether teachers with additional service or management responsibilities are more likely to participate in learning opportunities than are teachers who do not hold such positions.

METHODOLOGY

Participants and Procedure

The data were collected within the COACTIV study ("Professional Competence of Teachers, Cognitively Activating Instruction, and the Development of Students' Mathematical Literacy"; Kunter et al., 2007[1]), which was a part of the German extension to the 2003 cycle of OECD's Programme for International Student Assessment (PISA). The sample consisted of 1939 teachers (51.3% female) of mathematics, science and other subjects[2] (e.g., German, English, and physical education), who were drawn from a nationally representative sample of 198 German secondary schools. The teachers were recruited by the school principal, who also administered the questionnaire. All teachers participated voluntarily and remained anonymous throughout the entire study. Teachers' age ranged from 25 to 65 years (*M* = 47.4, *SD* = 9.4); teaching experience ranged from 1 to 44 years (*M* = 20.8, *SD* = 10.6). The association between teachers' age and length of teaching experience was *r* = .90, indicating that the two measures are almost interchangeable. Separate analyses were conducted with age and teaching experience as predictors, but the findings were equivalent. Therefore, in the following we report results for teachers' age as predictor.

Measures

Formal learning opportunities were measured by an open-ended question asking teachers to list all in-service training activities they had attended in the previous two years, including seminars, workshops, conferences and school-specific professional development meetings. Prompted by the instruction "Please enter all training activities you have attended since 2001 in the table below," teachers reported the topic of each activity as well as additional information (i.e., year of attendance, duration, subjective rating of effectiveness). This direct assessment of in-service training activities has the advantage that the responses provide a very specific and concrete representation of teachers' behaviour. A possible disadvantage of this assessment procedure may be that teachers have no record of their professional development activities and therefore cannot recall them all in detail. However, given the low average number of activities attended over the 2 year period (approx. 3), it seems likely that the teachers' recall is comprehensive and accurate. The number of in-service training activities was summed for each individual and used as indicator of participation in formal learning opportunities. The training activities were then grouped by content domain, based on a theoretical model of teacher competence developed in the COACTIV project (Baumert et al., 2009; Brunner et al., 2006; Krauss et al., 2004; Kunter et al., 2007). This model proposes that teacher knowledge, which is one aspect of teacher competence, can be decomposed into five areas: content knowledge, pedagogical content knowledge, pedagogical and psychological knowledge, organisational knowledge and counselling knowledge. This typology of knowledge provided the basis for the categorisation scheme used to classify the in-service training activities listed by participating teachers. The list of categories was extended during the classification procedure to account for topics that were not covered by the theoretical domains of teacher knowledge. The final categorisation scheme comprised the nine categories described in Table 1. The teachers listed a total of 5633 in-service training activities, which were classified independently by two trained raters. Interrater agreement was κ = .81 (Cohen, 1960). In cases of disagreement, the raters compared their codings and resolved discrepancies by discussion. The analyses were based on the total number of activities that teachers attended in a particular category.

Table 1. The nine categories of teachers' in-service training activities.

Category	*Description*	*Example*
1. Subject Content	Activities focusing on the content of a school subject, without explicit consideration of pedagogical aspects.	Stochastics, Geometry
2. Subject-Specific Pedagogy	Activities focusing on subject-specific instruction, including curricular and assessment-related activities.	New methods in mathematics, Open-ended tasks in mathematics instruction
3. Pedagogy and Psychology	Activities focusing on learning processes, classroom management, instructional strategies, the support of gifted children and the prevention, management, and diagnosis of psychopathologies and behavioural disorders.	Learning motivation, Preventing violence
4. School Organisation	Activities focusing on the goals, structure and development of the school, including all activities relating to school administration and school leadership.	School programme, Pedagogical school development
5. School System	Activities focusing on the school system as a whole, and not to individual schools in particular.	Educational law, Educational reform
6. Counselling	Activities focusing on the school system as a whole, and not to individual schools in particular.	Mediation, Working with parents
7. General Skills	Activities that are not solely geared to the teaching profession, covering topics such as technology	First aid, Internet basics
8. Teacher Licensing	Activities focusing on the attainment of additional teaching licenses. These activities have a long-term focus and comprise content-related and pedagogical domains.	Part-time degree in astronomy, Post-graduate qualification in computer science
9. Teacher Training	Activities in which teachers provided training for other teachers in content-related and pedagogical domains.	Train the trainer, Training for mentors of beginning mathematics teachers

Informal learning opportunities were assessed by two different indicators, namely teacher collaboration and the use of professional literature. Teacher collaboration measured how closely teachers cooperated with their colleagues in choosing instructional strategies, planning lessons and developing class materials (sample item: "How often do you discuss lesson content with your colleagues?", PISA-Konsortium Deutschland, 2006). Teachers rated the six items of the scale on a 4-point Likert scale ranging from (1) "never" to (4) "very often". The internal consistency of the scale was satisfactory with a Cronbach's alpha of .82. Use of professional literature was measured by an open-ended question asking teachers to estimate the number of hours per week they spent reading professional literature of any kind on an average week in an academic year (PISA-Konsortium Deutschland, 2006).

Work engagement and additional professional responsibilities. Teachers' work engagement was measured by four subscales from the Occupational Stress and Coping Inventory (AVEM, Klusmann, Kunter, Trautwein, Lüdtke, & Baumert, 2008; Schaarschmidt & Fischer, 1997). The subscales tapped subjective significance of work (sample item: "Work is my main focus in life"), career ambition ("I have high aspirations for my future career"), exertion ("I spare no effort at work") and perfectionism ("I always want my work to be faultless"). Teachers rated each item on a 5-point Likert scale ranging from (1) "strongly disagree" to (5) "strongly agree". The internal consistency of the scale was satisfactory with a Cronbach's alpha of .75. Teachers' additional professional responsibilities were assessed by an open-ended question ("Do you have additional responsibilities in your school? If yes, please specify them.") and subsequently categorised into management and service responsibilities. Teachers with management responsibilities hold leadership roles (e.g., principals, vice principals, department heads). Service responsibilities include tasks such as running the school library, school security, guidance counselling and technology support. The data were coded as two dummy variables: management responsibilities (0 = no, 1 = yes) and service responsibilities (0 = no, 1 = yes). Note that German teachers with any of these responsibilities may have a reduced teaching load, but that all of them – even principals – continue classroom teaching.

Analyses

Many statistical models assume linear relationships among variables. In our study, however, we were particularly interested in determining the type of relationship between teachers' age and professional development activities. We therefore needed a model that allowed us to explore linear and nonlinear relationships. One appropriate statistical method is polynomial regression analysis, which is a special case of multiple regression analysis (Kutner, Nachtsheim, & Neter, 2004). This model uses power functions of predictors (x, x^2, etc.) in a regression equation to estimate curvilinear relationships between predictor and dependent variables. The shape of the function can be determined by testing the significance of each predictor in the model (x = linear relationship, x^2 = quadratic relationship, etc.). In the present analyses, we addressed our hypotheses by considering a linear and quadratic term for age, as expressed by the following regression function:

$$Y_i = \beta_0 + \beta_1 \mathrm{Age}_\mathrm{i} + \beta_2 \mathrm{Age}_\mathrm{i}^2 + \varepsilon_i$$

where i = 1...n individuals.

The first and second sets of analyses are based on a regression model that examines the effect of age without additional covariates. The third set of analyses extends the model and predicts the uptake of formal and informal learning opportunities by teachers' work engagement and additional responsibilities. The level of significance was specified as $\alpha = .05$.

Table 2. Predicting the uptake of formal an informal learning opportunities by teacher's age.

	Model 1		*Model 2*	
	B (SE)	*β*	*B (SE)*	*β*
In-service training				
Age	−.04 (.01)	−.10*	−.05 (.01)	−.12*
Age²			−.00 (.00)	−.11*
*R*²		.01		.02
Teacher collaboration				
Age	−.00 (.00)	−.09*	−.00 (.00)	−.08*
Age²			.00 (.00)	.01
*R*²		.01		.01
Use of professional literature				
Age	.01 (.01)	.06*	.02 (.01)	.07*
Age²			.00 (.00)	.01
*R*²		.00		.00

Note. B = unstandardised regression coefficient, *SE* = standard error of unstandardised regression coefficient, *β* = standardised regression coefficient, *R*² = variance explained by the model. *$p < .05$.

RESULTS

Research Question 1: Does Teachers' Use of Formal and Informal Learning Opportunities Change Across the Career?

We addressed our first research question by estimating a set of polynomial regression models predicting the uptake of in-service training,[3] teacher collaboration and use of professional literature. The results of these analyses are shown in Table 2. There was a negative linear effect of age on uptake of in-service training activities in Model 1

(β_{age} = -.10, $p < .05$) and an additional negative quadratic effect in Model 2 (β_{age} = -.12, $p < .05$; $\beta_{age}{}^{2}$ = -.11, $p < .05$). In other words, teachers' uptake of in-service training across the career was represented by a quadratic function, beginning at a low level at the start of the career, reaching a peak in mid career and decreasing thereafter (see Panel A of Fig. 1). More specifically, teachers aged 27 participated on average in 2.89 in service courses in the 2 years period surveyed. The average participation rate increased to 3.72 courses at age 42 before decreasing again to 1.58 courses at age 65. In contrast, there was a negative linear effect of age on teacher collaboration in Model 1 (β_{age} = -.09, $p < .05$), but no additional quadratic effect in Model 2. In other words, teacher collaboration follows a linear pattern, with older teachers collaborating less frequently than younger teachers (see panel B of Fig. 1).

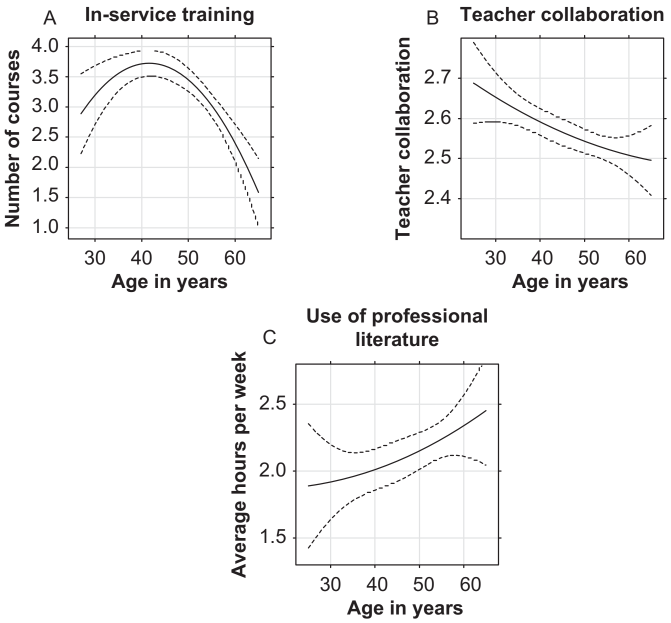

Figure 1. Uptake of formal and informal learning opportunities as a function of teachers' age.

Note. The solid line indicates the function predicted by the regression analysis and the dashed lines delimit the 95% confidence interval.

Finally, the models predicting the use of professional literature revealed a positive linear effect in Model 1 (β_{age} = .06, $p < .05$) but again no significant quadratic effect in Model 2. In other words, older teachers used professional literature more frequently than younger teachers (see panel C of Fig. 1). Estimated average reading time increased from 1.84 h per week for beginning teachers (age 27) to 2.38 h per week for teachers approaching retirement (age 65).

Research Question 2: Does Teachers' Use of Content-Specific In-Service Training Activities Change Across the Career?

The descriptive statistics in Table 3 provide an overview of the in-service training activities attended over the 2-year period surveyed. The data show that teachers predominantly attended activities on subject content (*M* = .70 courses in 2 years, *SD* = 1.22), subject-specific pedagogy (*M* = .74, *SD* = 1.26) and pedagogy and psychology (*M* = .65, *SD* = 1.06). Participation in activities targeting general skills (*M* = .40, *SD* = .81) and school organisation (*M* = .31 teacher attended in-service training on the school system (*M* = .09, *SD* = .43) or counselling (*M* = .10, *SD* = .38). Only a small group of teachers pursued training in the categories teacher licensing (*M* = .02, *SD* = .17) or teacher training (*M* = .02, *SD* = .21). These two categories were therefore excluded from our further analyses. Courses that could not be classified to one of the previous categories were combined in the category "other" (*M* = .17, *SD* = .17). This category accounted for only 5.3% of all training activities and was therefore not investigated further.

Table 3. Participation in formal learning opportunities by category.

Categories	*N*	*Min*	*Max*	*M*	*SD*
1. Subject Content	1759	0	13	0.70	1.22
2. Subject-Specific Pedagogy	1759	0	11	0.74	1.26
3. Pedagogy and Psychology	1759	0	7	0.65	1.06
4. School Organisation	1759	0	8	0.31	0.82
5. School System	1759	0	6	0.09	0.43
6. Counselling	1759	0	5	0.10	0.38
7. General Skills	1759	0	6	0.40	0.81
8. Teacher Licensing	1759	0	3	0.02	0.17
9. Teacher Training	1759	0	4	0.02	0.21
10. Other	1759	0	7	0.17	0.54

Note. Min and Max indicate the minimum and the maximum number of courses attended by participating teachers over the 2-year survey period.

Next, we predicted uptake of content-specific in-service training activities by teachers' age. We restricted these analyses to the seven largest categories because the others were not large enough to provide stable estimates. The results presented in Table 4 reveal significant change in four of the seven categories.

Table 4. Predicting the uptake of formal learning opportunities of different categories by teacher's age.

Categories	*Model 1*		*Model 2*	
	B (SE)	*β*	*B (SE)*	*β*
Subject Content				
Age	−.01 (.00)	−.07*	−.01 (.00)	−.09*
Age²			.00 (.00)	−.07*
*R*²		.01		.01
Subject-Specific Pedagogy				
Age	−.01 (.00)	−.08*	−.01 (.00)	−.09*
Age²			.00 (.00)	−.07*
*R*²		.01		.01
Pedagogy and Psychology				
Age	−.01 (.00)	−.08*	−.01 (.00)	−.09*
Age²			.00 (.00)	−.05*
*R*²		.01		.01
School Organisation				
Age	.00 (.00)	.03	.00 (.00)	.02
Age²			.00 (.00)	−.02
*R*²		.00		.00
School System				
Age	.00 (.00)	−.02	.00 (.00)	−.03
Age²			.00 (.00)	−.04
*R*²		.00		.00
Counselling				
Age	.00 (.00)	−.01	.00 (.00)	−.01
Age²			.00 (.00)	.01
*R*²		.00		.00
General Skills				
Age	.00 (.00)	−.03	.00 (.00)	−.05
Age²			.00 (.00)	−.08*
*R*²		.00		.01

Note. B = unstandardised regression coefficient, *SE* = standard error of unstandardised regression coefficient, *β* = standardised regression coefficient, R^2 = variance explained by the model.
$^*p < .05$.

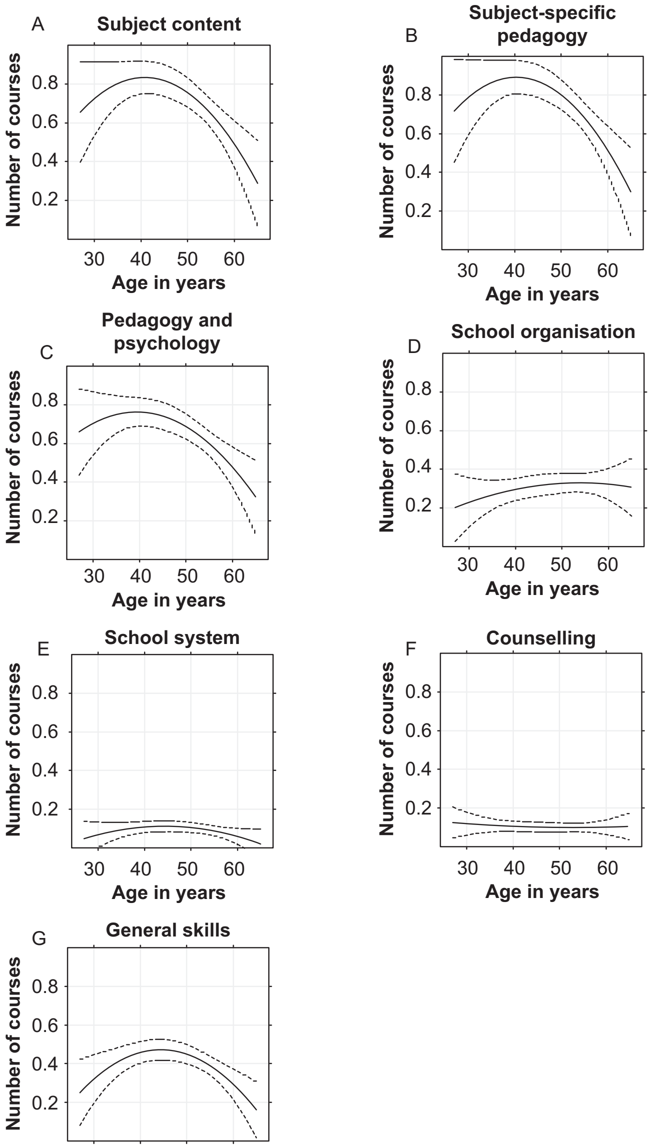

Figure 2. Uptake of formal learning opportunities of different categories as a function of teacher's age.

Note. The solid line indicates the function predicted by the regression analysis and the dashed lines delimit the 95% confidence interval.

Similar to the overall trend observed for in-service training activities, we found linear and quadratic effects for courses on subject content (Model 1: β_{age} = -.07, $p < .05$; Model 2: β_{age} = -.09, $p < .05$; β^2_{age} = -.07, $p < .05$), subject-specific pedagogy (Model 1: $\beta_{age\,=}$ -.08, $p < .05$; Model 2: β_{age} = -.09, $p < .05$; β^2_{age} = -.07, $p < .05$) and pedagogy and psychology (Model 1: β_{age} = -.08, $p < .05$; Model 2: β_{age} = -.09, $p < .05$; β^2_{age} -.05, $p < .05$). Participation in training activities targeting general skills followed a quadratic trend only (Model 2: β^2_{age} = -.08, $p < .05$). The linear and quadratic terms in all other models (school organisation, school system and counselling) were not significantly different from zero, which suggests that participation in these courses did not differ with teachers' age.

To facilitate the interpretation of the statistical effects, we plotted the estimated response functions in Fig. 2. As the graphs show, the peaks in panels A, B, C and G are located in the same age period, namely the age range of 38-45 years. These content- specific patterns are very similar to the function found for in- service training in general in the first set of analyses (Panel A of Fig. 1). This implies that the pattern observed for in-service training in general is produced primarily by teachers' uptake of training in these four areas.

Table 5. Bivariate correlations between the uptake of formal and informal learning opportunities and predictor variables.

	In-service trainig	*Teacher collaboration*	*Use of professional literature*
Age	−.11*	−.09*	.06*
Age^2	−.08*	.04*	.00
Gender[a]	.12*	−.18*	−.07*
Marital status[b]	−.06*	−.02	.02
Work engagement	.21*	.15*	.11*
Service responsibilities[c]	.07*	−.04	.06*
Management responsibilities[d]	.11*	.02	.03

$*p < .05$.

[a] Gender: 0 = Male; 1 = Female.

[b] Marital status: 0 = Living with a partner; 1 = Living alone.

[c] Service responsibilities: 0 = No service responsibilities in school; 1 = Service responsibilities in school.

[d] Management responsibilities: 0 = No management responsibilities in school ; 1 = Management responsibilities in school.

Research Question 3: Which Individual Characteristics Explain Age-Related Changes?

In the third set of analyses, we investigated whether the age patterns identified in the first part of our study are explained by teachers' work engagement and additional responsibilities. Prior to the analyses, we examined bivariate relationships between the predictor variables and uptake of the three types of learning opportunity (Table 5). In addition to linear and quadratic age effects, we found that gender was significantly correlated with uptake of all three types of learning opportunity, with females being more actively involved than males in in-service training ($r = .12, p < .05$) and teacher collaboration ($r = .18, p < .05$). However, females spent less time reading professional literature ($r = -.07, p < .05$). Further, teachers who lived alone pursued less in-service training than did teachers who lived with a partner ($r = -.06, p < .05$). Teachers' work engagement was also positively associated with the uptake of all three learning opportunities ($.11 < r < .21, p < .05$). Teachers with service responsibilities attended more in-service training activities ($r = .07, p < .05$) and read more professional literature ($r = .06, p < .05$); teachers with management responsibilities attended more in-service training activities ($r = .11, p < .05$).

Next, we examined whether the additional predictor variables explained uptake of learning opportunities and reduced the age effects identified in the first set of analyses. For this purpose, we fitted two regression models each for uptake of in-service training, teacher collaboration and use of professional literature. The first model included the demographic variables age, gender and marital status; the second model further included work engagement and additional responsibilities. The results presented in Table 6 show that demographic and work-related variables significantly predicted uptake of in service training. Model 1 shows effects of gender ($\beta = .10, p < .05$), marital status ($\beta = -.08, p < .05$) and age ($\beta_{age} = -.11, p < .05$; $\beta_{age} = -.09, p < .05$). The second model confirmed these demographic effects and further revealed that work engagement ($\beta = .18, p < .05$) service responsibilities ($\beta = .08, p < .05$) and management responsibilities ($\beta = .11, p < .05$) positively predicted uptake of in-service training. We also tested interaction terms between both (1) age and work engagement and (2) age and service/management responsibilities, but none were significant. The findings demonstrate that, even when demographic characteristics were controlled, teachers with high levels of work engagement and service or management responsibilities pursued more in-service training. The regression models predicting teacher collaboration revealed significant effects of age ($\beta = -.05, p < .05$) and gender ($\beta = .18, p < .05$) in Model 1, and a positive effect of work engagement ($\beta = .13, p < .05$) but no effects of additional responsibilities in Model 2. We again tested interaction terms in Model 2. We found a significant negative effect for the interaction between age and work engagement ($\beta = -.05$, $p < .05$), suggesting that the predictive value of work engagement decreases with teachers' age. Finally, the regression models predicting use of professional literature revealed significant effects for age ($\beta = .05, p < .05$) and gender ($\beta = -.07, p < .05$) in

Table 6. Predicting the uptake of formal and informal learning opportunities by individual characteristics.

	Model 1		*Model 2*	
	B (SE)	*β*	*B (SE)*	*β*
In-service training				
Age	−.04 (.01)	−.11*	−.04 (.01)	−.10*
Age^2	−.00 (.00)	−.09*	−.00 (.00)	−.08*
Gender[a]	.62 (.15)	.10*	.61 (.16)	.10*
Marital status[b]	−.70 (.22)	−.08*	−.61 (.22)	−.07*
Work engagement			.55 (.08)	.18*
Service responsibilities[c]			.62 (.19)	.08*
Management responsibilities[d]			.81 (.18)	.11*
R^2		.03		.09
Teacher Collaboration				
Age	.00 (.00)	−.05*	.00 (.00)	−.04
Gender[a]	.18 (.03)	.18*	.17 (.03)	.17*
Marital status[b]	−.06 (.03)	−.04	−.05 (.03)	−.04
Work engagement			.06 (.01)	.13*
Service responsibilities[c]			−.04 (.03)	−.03
Management responsibilities[d]			.02 (.03)	.02
R^2		.04		.05
Use of professional literature				
Age	.01 (.01)	.05*	.01 (.01)	.06*
Gender[a]	-.29 (.11)	−.07*	−.31 (.11)	−.07*
Marital status[b]	.21 (.16)	.03	.24 (.15)	.04
Work engagement			.25 (.06)	.12*
Service responsibilities[c]			.27 (.14)	.05
Management responsibilities[d]			.05 (.13)	.01
R^2		.01		.03

Note. B = unstandardised regression coefficient, *SE* = standard error of unstandardised regression coefficient, *β* = standardised regression coefficient, *R*^{2} = variance explained by the model.

$*p < .05$.

[a] Gender: 0 = Male; 1 = Female.

[b] Marital status: 0 = Living with a partner; 1 = Living alone.

[c] Service responsibilities: 0 = No service responsibilities in school; 1 = Service responsibilities in school.

[d] Management responsibilities: 0 = No management responsibilities in school ; 1 = Management responsibilities in school.

Model 1. Model 2 confirmed these effects and revealed an additional positive effect for work engagement (β = .12, $p < .05$), but no significant effects for additional service or management responsibilities. Interaction terms were again tested but found not to be significant.

SUMMARY AND DISCUSSION

We used Huberman's (1989) teacher career stage model as a theoretical framework to derive empirically testable hypotheses about the uptake of learning opportunities across the teaching career. The study was based on a large sample of mathematics teachers of different age groups. This heterogeneity of the sample allowed us to approximate teachers' uptake of learning opportunities across the entire career cycle. The learning opportunities distinguished were formal training activities and two examples of informal learning. The first two parts of the study investigated the quantity and the quality of the learning opportunities attended from a lifespanperspective. The third part examined individual predictors of age-related differences. More specifically, we tested whether work engagement and additional responsibilities explained the changes observed across the teaching career.

Uptake of Learning Opportunities From a Lifespan Perspective

We found distinct trajectories of teachers' uptake of the three types of learning opportunities across the career. Uptake of in- service training described a curvilinear pattern, with the highest uptake in mid-career (around age 42), confirming the hypothesis that teachers pursue formal learning opportunities primarily during the phase of experimentation and activism. In contrast, the SASS (1999-2000) data indicated constant participation in in-service training across age groups (Choy et al., 2006). How can this difference be explained? In Germany, teachers are not required to attend professional development training to renew their teaching license, and the state generally does not specify the number of courses that teachers are required to attend (Avenarius & Heckel, 2000). In contrast, U.S. teachers are generally obliged to attend in-service courses on a regular basis to meet the requirements of their state (NASDTEC, 2004). Our data therefore reflect largely voluntary participation, whereas the U.S. data reflect the consequences of strict participation requirements. Teacher collaboration, in contrast, decreased linearly over the career. This finding deviates slightly from our hypothesis of a quadratic trend, as indicated by the SASS (1999-2000) data. The finding that teachers collaborate more at the beginning of their career than in the middle or at the end may be attributable to younger teachers still being more eager to learn from and draw on the professional expertise of more experienced teachers (Grangeat & Gray, 2007). We expected teachers to spend less time reading professional literature as they approached retirement. However, the data showed that older teachers spent more time reading than their younger colleagues. This finding suggests that older teachers do not invest less time in professional development than their younger peers, but that they prefer different media or learning opportunities. Alternatively, it can be hypothesised that self-directed learning is more attractive to older teachers, who therefore choose professional literature as their means for learning.

After investigating the different types of learning opportunities, we next focused on the content of the in-service training activities. Through its specific assessment of the topics covered in each training activity, our study went beyond quantitative characteristics and analyzed the heterogeneity of all training activities attended. Our study design thus went beyond the scope of the SASS (1999-2000), which measured the content of professional development activities in predefined categories. The results indicated statistically significant age-related change in the uptake of courses in the categories subject content, subject-specific pedagogy, pedagogy and psychology and general skills. The changes followed a curvilinear pattern which peaked similarly to the overall trend for in-service training din the middle of the career. Therefore, the data do not provide evidence that the different categories are particularly relevant at different phases of the career. Again, this result differs from findings based on SASS (1999-2000) data, which demonstrated that highly experienced teachers (>20 years) were more likely to attend courses on student assessment and computers, whereas younger teachers (1-3 years of experience) were more likely to attend courses on classroom management. These contrasting results may again be attributable to differences in professional development requirements. Older U.S. teachers are required to participate in professional development activities, and therefore choose the courses that are most beneficial to them at their career stage. Teachers in Germany do not have the same external incentives to participate in formal courses at the end of their career.

Individual Predictors of the Uptake of Learning Opportunities

We further investigated teachers' work engagement and additional responsibilities and examined their relationship to professional learning. Findings demonstrated that teachers with high work engagement and teachers who held service or management responsibilities pursued more in-service training. Further, teachers with high work engagement but not teachers with service or management responsibilities used more informal learning opportunities. Moreover, our data showed that the age effects were not fully explained by the inclusion of the additional individual predictors. This suggests that the age-related patterns observed cannot be attributed to an overall decrease in engagement, as at least partially implied by the career stage model. Future studies need to examine alternative explanations for the remaining differences across the career. Industrial and organisational psychologists have examined professional development behaviour in other professions (Maurer, Weiss, & Barbeite, 2003; Simpson, Greller, & Stroh, 2002; Staudinger & Baumert, 2007; Warr & Birdi, 1998). All of these studies have investigated linear change across time, meaning that the scope to draw comparisons with our findings is limited. However, the results showed that older employees participated less frequently in professional development than did younger employees. These findings are consistent with ours and suggest that the mechanisms that explain the uptake of learning opportunities may be generalisable across professions. If this is indeed the case, it would be helpful to draw on generic, non-teacher-specific theories of lifespan development to explain changes over the career. Following the lifespan approach, behavioural changes (in our case, teachers' learning behaviour) may be attributable to an array of age-related changes in cognitive abilities, but also to motivational and volitional variables and to changes in the social context. One theory that may provide an interesting approach to describing teachers' learning behaviour is socio emotional selectivity theory, which explains reduced involvement in professional learning in terms of a reduced need for information and knowledge from mid-career onwards (Carstensen, Isaacowitz, & Charles, 1999). The theory suggests that individuals tend to prioritise short-term over long-term goals and to pursue less new information when they become aware of their limited time perspective. In the light of this theory, it could be argued that older teachers reduce their attendance of in-service training because the potential payoffs (e.g., promotion, instructional improvements) become smaller as they approach retirement.

Limitations and Implications

We now discuss some limitations of our study and indicate possible implications for research and policy. First, the cross- sectional study design restricts the interpretation of our findings. Our data provide insights into age-related differences but not into intra individual change. It is not possible to determine whether the patterns observed are due to cohort effects or reflect intra- individual differences. Longitudinal studies are therefore needed to track individual teachers' learning behaviour and professional development over an extended period of time. Second, our findings need to be interpreted with reference to the professional development requirements of the state in question (Eurydice, 2008; NASDTEC, 2004). These requirements influence whether and to what extent teachers participate in professional development. Because cross-state policy differences make it difficult to compare findings from different studies, future research needs to be explicit about the context in which teacher professional development takes place. Third, we acknowledge that this study took an individual perspective on professional development. Research on teacher behaviour has indicated that characteristics of the school or work context (i.e., principal support, material resources, etc.) can also impact teachers' work engagement (Klusmann et al., 2008), commitment (Firestone & Pennell, 1993) and professional development behaviour (Kwakman, 2003). Therefore, future studies need to examine whether and to what extent the school context impacts the development of teachers' learning behaviour across the career cycle. Finally, this study was based primarily on mathematics and science teachers at secondary schools. Therefore, the results cannot be generalised to other groups of teachers (e.g., elementary school teachers), who were not represented in the sample. In conclusion, this study has both theoretical implications for research on professional development and practical implications for policy makers. The study investigated teachers' participation in professional development from a new perspective, using a generic model of teachers' career development to make predictions about their actual behaviour. The model provided insights into teachers' changing needs and concerns, but its capacity to explain different developmental patterns of formal and informal professional development activities was limited. More specific theoretical models are thus needed that describe teachers' learning behaviour from a developmental perspective. The study's findings also have practical implications for policy makers who plan and provide professional development opportunities for teachers. This applies in particular to the in-service training provided by local and state agencies (i.e., teacher training institutes). Our results indicate that older teachers show reduced involvement in in-service training. Hence, efforts are needed to increase older teachers' participation and to promote lifelong teacher learning. Given the design of the present study, we cannot specify the steps to be taken by policy makers, but we can suggest three possible ways of increasing participation in professional development activities. First, policy makers could make it compulsory for teachers to complete a minimum number of in-service training hours within a given period of time (see also Eurydice, 2003), thus ensuring that all teachers participate in learning activities on a regular basis. Second, participation rates might be raised by offering activities that respond to the needs of older teachers. Courses on topics that were not covered during their university education (i.e., use of technology for instruction) may be more relevant for this group of teachers and thus increase their uptake of professional development opportunities. Finally, it might be helpful to provide opportunities for experienced teachers to share and learn from each other. This two-way process could be motivating for teachers and help them to respond more effectively to the needs of their school.

ACKNOWLEDGEMENT

We thank the members of the COACTIV team, Susannah Goss for language editing and Johanna Seiz for data coding.

NOTES

[*] Reprint from pp. 116-126, Teaching and Teacher Education, 27 (2011).

[1] The project was funded by the German Research Foundation (DFG; BA 1461/2-2) as part of its Priority Program on School Quality (BIQUA).

[2] A detailed list of subjects is not provided because every participating teacher is licensed for at least two subjects, resulting in a large amount of possible subject combinations.

[3] Teachers with less than 2 years of teaching experience (excluding teacher training) were excluded from the analyses of in-service training activities because they had not yet had 2 full years' opportunity to participate in such courses.

REFERENCES

Anderson, L., & Olsen, B. (2006). Investigating early career urban teachers' perspectives on and experiences in professional development. *Journal of Teacher Education, 57*(4), 359–377. doi:10.1177/0022487106291565

Avenarius, H., & Heckel, H. (2000). *Schulrechtskunde: ein Handbuch für Praxis, Rechtsprechung und Wissenschaft.* [*School legal studies: a handbook for practice, jurisprudence, and science*] (7th ed.). Neuwied: Luchterhand.

Ball, D. L., & Cohen, D. K. (1999). Developing practice, developing practitioners: toward a practice-based theory of professional education. In L. Darling-Hammond, & G. Sykes (Eds.), *Teaching as the learning profession: Handbook of policy and practice* (pp. 3–32). San Francisco, CA: Jossey-Bass.

Baumert et al. (2009). Teachers' mathematical knowledge, cognitive activation in the classroom, and student progress. *American Educational Research Journal*. doi:10.3102/0002831209345157

Baumert, J., & Kunter, M. (2006). Stichwort: Professionelle Kompetenz von Lehrkräften. [Teachers' professional competence]. *Zeitschrift für Erziehungswissenschaft, 9*(4), 469–520. doi:10.1007/s11618-006-0165-2

Brekelmans, M., Wubbels, T., & van Tartwijk, J. (2005). Teacher student relationships across the teaching career. International *Journal of Educational Research, 43*(1–2), 55–71. doi:10.1016/j.ijer.2006.03.006

Brunner et al. (2006). Die professionelle Kompetenz von Mathematiklehrkräften: conzeptualisierung, Erfassung und Bedeutung für den Unterricht: Eine Zwischenbilanz des COAC- TIV-Projektes. [The professional competencies of mathematics teachers: con- ceptualisation, assessment, and significance for instruction. An interim review of the COACTIV project]. In M. Prenzel, & L. Allolio-Näcke (Eds.), *Untersuchungen zur Bildungsqualität von Schule: Abschlussbericht des DFG-Schwerpunktprogramms* (pp. 54–82). Münster: Waxmann.

Carstensen, L. L., Isaacowitz, D. M., & Charles, S. T. (1999). Taking time seriously: a theory of socioemotional selectivity. *American Psychologist, 54*(3), 165–181. doi:10.1037/0003-066X.54.3.165

Choi, P. L., & Tang, S. Y. F. (2009). Teacher commitment trends: cases of Hong Kong teachers from 1997 to 2007. *Teaching and Teacher Education, 25*, 767–777.

Choy, S. P., Chen, X., & Bugarin, R. (2006). *Teacher professional development in 1999-2000: What teachers, principals and district staff report (NCES 2006–305)*. Washington, DC: National Center for Education Statistics.

Cohen, J. (1960). A coefficient of agreement for nominal scales. *Educational and Psychological Measurement, 20*(1), 37–46. doi:10.1177/001316446002000104

Cookson, P. S. (1986). A framework for theory and research on adult-education participation. *Adult Education Quarterly, 36*(3), 130–141. doi:10.1177/0001848186036003002

Corcoran, T. B. (2007). *Teaching matters: How state and local policymakers can improve the quality of teachers and teaching.* Philadelphia: Consortium for Policy Research in Education.

Dall'Alba, G., & Sandberg, J. (2006). Unveiling professional development: a critical review of stage models. *Review of Educational Research, 76*(3), 383–412. doi:10.3102/00346543076003383

Day, C., Sammons, P., Stobart, G., Kington, A., & Gu, Q. (2007). *Teachers matter: Connecting work, lives and effectiveness.* Berkshire: Open University Press.

Desimone, L. M. (2009). Improving impact studies of teachers' professional development: Toward better conceptualizations and measures. *Educational Researcher, 38*(3), 181–199. doi:10.3102/0013189X08331140

Desimone, L. M., Smith, T., & Ueno, K. (2006). Are teachers who need sustained, content-focused professional development getting it? An administrator's dilemma. *Educational Administration Quarterly, 42*(2), 179–215. doi:10.1177/0013161X04273848

Eurydice. (2002). *The teaching profession in Europe: Profile, trends and concerns.* Report I. Initial training and transition to working life of teachers in general lower secondary education. Brussels: Author.

Eurydice. (2003). *The teaching profession in Europe: Profile, trends and concerns*. Report III: Working conditions and pay. Brussels: Eurydice.

Eurydice. (2008). *Levels of autonomy and responsibilities of teachers in Europe*. Brussels: Author.

Feiman-Nemser, S. (2001). From preparation to practice: designing a continuum to strengthen and sustain teaching. *Teachers College Record, 103*(6), 1013–1055. doi:10.1111/0161-4681.00141

Fessler, R. (1995). Dynamics of teacher career stages. In T. R. Guskey, & M. Huberman (Eds.), *Professional development in education* (pp. 171–192). New York: Teachers College Press.

Firestone, W. A., & Pennell, J. R. (1993). Teacher commitment, working conditions, and differential incentive policies. *Review of Educational Research, 63*(4), 489–525. doi:10.3102/00346543063004489

Glazerman et al. (2008). *Impacts of comprehensive teacher induction: Results from the first year of a randomized controlled study (NCEE 2009–4034)*. Retrieved from http://ies.ed. gov/ncee/pdf/20094035.pdf

Goddard, Y. L., Goddard, R. D., & Tschannen-Moran, M. (2007). A theoretical and empirical investigation of teacher collaboration for school improvement and student achievement in public elementary schools. *The Teachers College Record, 109*(4), 877–896.

Grangeat, M., & Gray, P. (2007). Factors influencing teachers' professional competence development. *Journal of Vocational Education and Training, 59*(4), 485–501. doi:10.1080/13636820701650943

Gregorc, A. F. (1973). Developing plans for professional growth. *NASSP Bulletin, 57*, 1–8. doi:10.1177/019263657305737701

Hill, H. C. (2007). Learning in the teaching workforce. *The Future of Children, 17*(1), 111–127. doi:10.1353/foc.2007.0004

Huberman, M. (1989). The professional life cycles of teachers. *Teacher College Record, 91*(1), 31–58.

Klusmann, U., Kunter, M., Trautwein, U., Lüdtke, O., & Baumert, J. (2008). Engagement and emotional exhaustion in teachers: does school context make a difference? *Applied Psychology: An International Review, 57*, 127–151. doi:10.1111/j.1464-0597.2008.00358.x

Krauss et al. (2004). COACTIV: Professionswissen von Lehrkräften, kognitiv aktivierender Mathematikunterricht und die Entwicklung von mathematischer Kompetenz. [COACTIV: professional competence of teachers, cognitively activating instruction, and development of students' mathematical literacy]. In J. Doll, & M. Prenzel (Eds.), *Bildungsqualität von Schule: Lehrerprofessionalisierung, Unterrichtsentwicklung und Schülerforderung als Strategien der Qualitätsverbesserung* (pp. 31–53). Münster: Waxmann.

Kunter et al. (2007). Linking aspects of teacher competence to their instruction: results from the COACTIV project. In M. Prenzel (Ed.), *Studies on the educational quality of schools. The final report of the DFG priority programme* (pp. 32–52). Münster, Germany: Waxmann.

Kutner, M. H., Nachtsheim, C. J., & Neter, J. (2004). *Applied multiple regression models* (4th ed.). New York, NY: McGraw-Hill.

Kwakman, K. (2003). Factors affecting teachers' participation in professional learning activities. *Teaching and Teacher Education, 19*, 149–170. doi:10.1016/ S0742-051X(02)00101-4

Lieberman, A. (1995). Practices that support teacher development. *Phi Delta Kappan, 76*, 591–596.

Little, J. W. (1993). Teachers' professional development in a climate of educational reform. *Educational Evaluation and Policy Analysis, 11*(2), 129–151. doi:10.3102/01623737015002129

Lohman, M. C. (2000). Environmental inhibitors to informal learning in the work- place: A case study of public school teachers. *Adult Education Quarterly, 50*(2), 83–101. doi:10.1177/07417130022086928

Maurer, T. J., Weiss, E. M., & Barbeite, F. G. (2003). A model of involvement in work- related learning and development activity: the effects of individual, situational, motivational, and age variables. *Journal of Applied Psychology, 88*(4), 707–724. doi:10.1037/0021-9010.88.4.707

Mesler, L., & Spillane, J. P. (2009, April). *Teacher learning and instructional change: How formal and on-the-job learning opportunities predict change in elementary school teachers' practice*. Paper presented at the Annual Meeting of the American Educational Research Association, San Diego, CA.

Mok, Y. F., & Kwon, T. M. (1999). Discriminating participants and non-participants in continuing professional education: The case of teachers. *International Journal of Lifelong Education, 18*(6), 505–519. doi:10.1080/026013799293559

National Association of State Directors of Teacher Education and Certification (NASD-TEC). (2004). *Table E1: Professional development: general description*. Retrieved May 28, 2009, from https://www.nasdtec.info/View/TopicalTables.aspx

National Commission on Teaching and America's Future. (2003). *No dream denied: a pledge to America's children*. Washington: National Commission on Teaching and America's Future.

Peterson, W. (1964). Age, teacher's role and institutional setting. In B. Biddle, & W. Elena (Eds.), *Contemporary research on teacher effectiveness* (pp. 264–315). New York: Holt, Rinehart and Winston.

PISA-Konsortium Deutschland. (Ed.). (2006). *PISA 2003: Dokumentation der Erheb*ungsinstrumente [*PISA 2003: Documentation of assessment instruments*]. Münster: Waxmann.

Putnam, R. T., & Borko, H. (2000). What do new views of knowledge and thinking have to say about research on teacher learning? *Educational Researcher, 29*(1), 4–15. doi:10.3102/0013189X029001004

Schaarschmidt, U., & Fischer, A. W. (1997). AVEM - ein diagnostisches instrument zur differenzierung von typen gesundheitsrelevanten verhaltens und erlebens gegenüber der arbeit. [AVEM - an instrument for diagnosing different types of work- and health-related behavior and experience]. *Zeitschrift für Differentielle und Diagnostische Psychologie, 18*(3), 151–163.

Scribner, J. P. (1999). Professional development: Untangling the influence of work context on teacher learning. *Educational Administration Quarterly, 35*(2), 238–266. doi:10.1177/0013161X99352004

Sikes, P. J., Measor, L., & Woods, P. (1985). *Teacher careers: Crisis and continuities*. Lewes, UK: Falmer.

Simpson, P. A., Greller, M. M., & Stroh, L. K. (2002). Variations in human capital investment activity by age. *Journal of Vocational Behavior, 61*(1), 109–138. doi:10.1006/jvbe.2001.1847

Staudinger, U. M., & Baumert, J. (2007). Bildung und Lernen jenseits der 50: plastizität und Realität. [Education and learning beyond 50: plasticity and reality]. In P. Gruss (Ed.), *Die Zukunft des Alterns. Die Antwort der Wissenschaft* (pp. 240–257). Munich: Beck.

U. S. Congress. (2001). *No Child Left behind Act of 2001: Public Law 107-110, 107th Congress*. Washington, DC: Government Printing Office.

U.S. Department of Education, & NationalCenter for Education Statistics. (2002). *Schools and Staffing Survey, 1999-2000: Overview of the data for public, private, public charter, and Bureau of Indian Affairs elementary and secondary schools, NCES 2002-313*. Washington, DC: U.S. Department of Education, & National Center for Education Statistics.

Unruh, A., & Turner, H. (1970). Supervision for change and innovation. Boston: Houghton Mifflin.

Veenman, S. (1984). Perceived problems of beginning teachers. *Review of Educational Research, 54*(2), 143–178. doi:10.3102/00346543054002143

Warr, P., & Birdi, K. (1998). Employee age and voluntary development activity. *International Journal of Training and Development, 2*(3), 190–204. doi:10.1111/1468-2419.00047

AFFILIATIONS

Dirk Richter
Max Planck Institute for Human Development

Mareike Kunter
Max Planck Institute for Human Development

Uta Klusmann
Max Planck Institute for Human Development

Oliver Lüdtke
Max Planck Institute for Human Development

Jürgen Baumert
Max Planck Institute for Human Development

RAINER BROMME, STEPHANIE PIESCHL & ELMAR STAHL

EPISTEMOLOGICAL BELIEFS AND STUDENTS' ADAPTIVE PERCEPTION OF TASK COMPLEXITY

INTRODUCTION

Epistemological Beliefs

Epistemological beliefs are usually defined as beliefs about knowledge and knowing. One of the most widely used framework within educational psychology (Buehl & Alexander, 2001; Hofer & Pintrich, 1997), among others widely used (Niessen, Vermunt, Abma, Widdershoven, & van der Vleuten, 2004), comprises four identifiable and more or less interrelated dimensions of beliefs. According to Hofer and Pintrich (1997) the first two dimensions represent the "nature of knowledge": (a) the certainty of knowledge is focused on the perceived stability and the strength of supporting evidence, and (b) the structure of knowledge describes the relative connectedness of knowledge. The other two dimensions describe the nature of "knowing": (a) the justification of knowledge explains how individuals proceed to evaluate and warrant knowledge claims, and (b) the source of knowledge describes where knowledge resides, internally and/or externally. In the remainder of this paper we will primarily refer to this framework, although, from the beginning alternative frameworks have been proposed (for an overview with an emphasis on the assumed dimensions, see Buehl, 2008). More recently, Greene, Azevedo, and Torney-Purta (2008), for example, proposed an alternative framework in which "justification", either personal or by authorities, was proposed as the core epistemic dimension, whereas beliefs about the simplicity and certainty of knowledge are coined "ontological" because they refer to learners' assumptions about the structure of the categorical representation of the world. In a similar vein, Bromme, Kienhues, and Stahl (2008) have argued that epistemological judgments are based on topic- and domain-related ontological assumptions. Both propositions point to the interplay between epistemological beliefs and topic- and domain-related knowledge, because ontological assumptions are very abstract knowledge about a domain. In the present study the interplay between epistemological beliefs and domain-specific knowledge is investigated.

An important assumption in epistemological theories is that learners' epistemological beliefs develop (or should develop) from the so-called "naïve" to "sophisticated" epistemologies (Hofer & Pintrich, 1997). The term "naïve" is used, for example, to indicate a person's belief that the knowledge to be learned consists

S. Krolak-Schwerdt et al., (Eds.), Teachers' Professional Development, 123–151.

of a stock of certain facts related to each other additively and whose veracity is guaranteed by an authority. Such facts, once found, mirror the world unambiguously. Through (formal) education people become aware that knowledge is, for example, more complex and relativistic thus resulting in a focus on the evaluation of different viewpoints (King & Kitchener, 2002). Persons with sophisticated perspective believe, for example, that the veracity of knowledge claims depends on context and is continuously established within social interactions, and that knowledge is rather a complex network of facts, theories and conjectures than a pure addition of true facts. They accept uncertainty and changeability of truth and the notion that knowledge is construed rather than given; however, this does not mean that it would be reasonable to conceive each knowledge claim in each context that way, for example, to doubt that the earth is (nearly) round. On the contrary, sophistication entails adaptability to contextual demands (Bromme, Kienhues, & Porsch, 2008; Elby & Hammer 2001). In the present study students' adaptivity to differences between learning tasks when planning their use of a complex learning environment was investigated. Thereby, the assumed relationship between epistemological beliefs, domain-specific knowledge, and adaptive planning behavior was scrutinized.

An increasing number of empirical studies show that more sophisticated epistemological beliefs are related to more adequate learning strategies and better learning outcomes. For example, students' epistemological beliefs have been found to influence their processing of information (Schommer, 1990), their academic performance (Schommer, 1993), their conceptual change (Mason & Boscolo, 2004), their quality of argumentation (Weinstock & Cronin, 2003), and their engagement in learning (Hofer & Pintrich, 1997). Although there are fewer studies concerning computer-based learning environments their results are encouraging as well. For example, Jacobson and Spiro (1995) found that learners with sophisticated epistemological beliefs were better able to learn and apply their knowledge after using a hypertext system than students with naïve epistemological beliefs. Additionally, epistemological beliefs were a good predictor of learning outcomes during hypertext learning (Bendixen & Hartley, 2003; Windschitl & Andre, 1998). There is also evidence that epistemological beliefs affect students' information retrieval from the Internet (Bråten, Strømsø, & Samuelstuen, 2005; Hofer, 2004), especially in more open-ended tasks (Tu, Shih, & Tsai, 2008), and the understanding of multiple documents (Strømsø, Bråten, &Samuelstuen, 2008). With regard to *self-regulated learning*, epistemological beliefs have also been related to the use of more self-reported (Cano, 2005; Dahl, Bals, & Turi, 2005; Neber & Schommer-Aikins, 2002) as well as concurrently measured (Kardash & Howell, 2000) metacognitive strategies, better metacognitive comprehension monitoring (Schommer, 1990), and more metacognitively controlled help-seeking in a hypertext (Bartholomé, Stahl, Pieschl, & Bromme, 2006).

How are epistemological beliefs related to metacognition? Most recent views on epistemological beliefs and learning conceive such beliefs as being involved in metacognitive processes of monitoring (Kuhn, 2000). For example, Kitchener

(1983) proposed three levels of cognitive processing: (a) *cognition* (all cognitive operations such as reading, memorizing, perceiving, computing), (b) *metacognition* (all cognitions that have cognitive operations as their subjects, for example comprehension monitoring), and (c) *epistemic cognition* (cognitions about the limits of knowing, the certainty of knowledge, or the criteria for knowledge). Only the epistemic cognitions are assumed to be involved in monitoring the epistemic nature of problems and the evaluation of solutions. To give another example, Hofer (2004) details how epistemological belief dimensions can be matched to components of metacognition. Beliefs about the nature of knowledge (certainty of knowledge and simplicity of knowledge) are assumed to share similarities with declarative metacognitive knowledge (Schraw & Moshman, 1995). Beliefs about the nature of knowing on the other hand (source and justification of knowledge) can be matched to the procedural component of metacognition, for example metacognitive monitoring (Schraw & Moshman, 1995). These models, concerned with structural aspects of epistemological beliefs (where they are located in the cognitive system), are promising but more functional theories about the impact of epistemological beliefs are rare (i.e., how do they exactly exert their influence?).

The COPES Model

An encouraging theoretical framework that helps to specify such a functional relationship is given by the COPES model of self-regulated learning (Greene & Azevedo, 2007; Muis, 2007; Winne & Hadwin, 1998): Epistemological beliefs are conceptualized as important internal conditions for learning, which impact learners' internal standards for metacognitive monitoring and control and, thereby, influence the whole learning process. More specifically, self-regulated learning according to the COPES model occurs in four weakly sequenced and recursive stages: (a) task definition, (b) goal setting and planning, (c) enactment, and (d) adaptation. In the task definition stage, a student generates her or his own perception about what the studying task is and what constraints and resources are in place. An important product of this stage is the student's perception of the given goal of the task. Based on this perception the student generates idiosyncratic goal(s) and constructs a plan for addressing that study task in the second stage. In the enactment stage the previously created plan of study tactics is carried out. The adaptation stage pertains to fine-tuning of strategies within the actual learning task as well as to long-term adaptations.

All four stages are embedded in the same general cognitive architecture. In the centre of this architecture are processes of metacognitive monitoring and control that students are assumed to use to self-regulate their learning process according to perceived task demands. If and how these processes occur depends on five constituents whose acronym gave the model its name, namely conditions (C), operations (O), products (P), evaluations (E) and standards (S). Conditions (C) pertain to external task conditions (e.g., task complexity) as well as to internal conditions (e.g., epistemological beliefs) and are assumed to directly influence learners' internal standards and their operations. Operations (O) include all cognitive processes that learners utilize to solve a learning task and which create internal or external products (P) (e.g., written answers). Students' goals are represented as multivariate profiles of standards (S) (e.g., targeted level of understanding). As a result of metacognitive monitoring, evaluations (E) are generated based on a comparison of students' products and standards. When a learner notices discrepancies she or he is able to perform metacognitive control by executing fix-up operations. To summarize, the COPES model describes how students might adapt their self-regulated learning process to important external conditions such as task complexity. Furthermore, it specifies the impact of learner-related internal conditions such as epistemological beliefs.

The Present Study

Based on the COPES model but with a special focus on the impact of epistemological beliefs on adaptivity to task complexity, the present study explores if epistemological beliefs affect processes of *metacognitive calibration.*

Traditionally, calibration refers to the accuracy of a person's subjective metacognitive judgments (e.g., their judgments of learning (JOL) regarding the confidence in recall) regarding their own objective performance (e.g., in a recall task of paired associates such as "ocean-tree"; example taken from Nelson & Dunlosky, 1991). Multiple measures of accuracy have been suggested in the literature on traditional calibration: The most frequently used method is *relative calibration* (Nelson, 1996) which denotes the degree of association between judgments and performance (e.g., the Goodman-Kruskal's Gamma correlation used by Nelson & Dunlosky, 1991). Additionally, indices of *absolute calibration* are often computed that indicate the exact degree of over- or underconfidence of judgments in relation to performance (e.g., bias score, see Schraw, 1995). Furthermore, measures of *discrimination* denote the ability to discriminate between the occurrence and the nonoccurrence of an event, for example predict correct versus incorrect performance (see Weingardt, Leonesio, & Loftus, 1994).

For the present study, we transferred the methodology of traditional calibration research outlined above to a new context (for more detail see Pieschl, 2009): We define metacognitive calibration as the alignment between learners' subjective *task definitions, goals and plans* (measured by the COPES questionnaire, see below) and objective task demands, more specifically *task complexity* (operationalized by Bloom's revised taxonomy; Anderson et al., 2001, see below). Therefore, metacognitive calibration in this sense denotes the degree of adaptivity to task complexity. Note that our definition of calibration is conceptually different from the traditional one, but that the same methods are applied: We assume that students *discriminate* between tasks of different complexity by indicating different task definitions, goals and plans. And we assume that students systematically *calibrate* their task definitions, goals and plans to task complexity, for example by planning more elaborate learning strategies for more complex tasks.

Assuming that epistemological beliefs affect metacognitive calibration as outlined above implies not only main effects of epistemological beliefs but also potential interactions between epistemological beliefs and task complexity. To illustrate the potential *main effects* of epistemological beliefs imagine a learner with a naïve belief that knowledge is simple and stable. As epistemological beliefs are assumed to directly influence the learner's internal standards, the learner might set quite superficial goals (e.g., "The goal is achieved if I have memorized the facts"; "I will complete this task in a short time") compared to a more sophisticated learner who believes that knowledge is complex and relative (e.g., "I have to deeply understand the subject-matter in order to apply it"; "I will need much time to complete the task"). To give another example, epistemological beliefs are also assumed to directly influence the learner's operations; thus, a more naïve learner might also plan rather superficial learning tactics and strategies for task completion (e.g., memorizing) compared to a more sophisticated learner who might plan strategies of deeper elaboration (e.g., critically evaluating).

To illustrate potential *interactions* with task complexity, consider that such differences might become more pronounced for more complex tasks. Specifically, if learners are confronted, for example, with the complex task of writing a pro- and contra- argumentation about a controversial topic, this task might be interpreted in multiple ways. A student who believes that knowledge is uncertain (sophisticated belief) would probably plan to verify each argument by searching for additional information, whereas a more naïve student would probably take each argument at face value. For a very simple task like a factual question on the other hand, these potential differences should be smaller, that is, students with naïve beliefs are assumed to approach such task superficially because they are assumed to have a general bias to underestimate task complexity and thus might approach all tasks too superficially. On the other hand, students with sophisticated beliefs should be able to accurately diagnose task demands and thus also plan adequately superficial strategies. Therefore, it was hypothesized that students with more sophisticated epistemological beliefs should show a better fit between external task demands such as task complexity and their self-regulated learning process.

Within a large project (for an overview: Bromme, Pieschl, & Stahl, 2010; Pieschl, Stahl, & Bromme, 2013) this general assumption about epistemological beliefs as important predictor of metacognitive calibration was tested with regard to each stage of self-regulated learning as defined in the COPES model. The present study – as well as an exploratory study already conducted (Stahl, Pieschl, & Bromme, 2006) – focused on the first two preparatory stages of self-regulated learning, that is, on COPES' *task definition* and *goal setting and planning*. All studies within this large project have common elements, that is, students work with (or plan to work with) a hypermedia information system on "genetic fingerprinting". This topic was chosen because it was judged inherently interesting by students, there was sufficient variance in students' epistemological beliefs towards this topic, and because this domain contains certain facts (e.g., "DNA contains four bases: adenine, cytosine, guanine, and thymine.") as well as controversial issues (e.g., "Should we compile comprehensive data bases of DNA profiles?"). Additionally, in all studies within this large project students have to complete (or plan to complete) learning tasks with different levels of complexity according to Bloom's revised taxonomy (Anderson et al., 2001). In this theoretical framework, task complexity is not defined quantitatively (e.g., by the number of necessary operations), but rather by the quality of the required cognitive processes. For example, factual questions, such as "What is the capital of Germany?" (correct answer: Berlin), are always considered simple because they just require recall of information from long-term memory. This even holds if task difficulty (i.e., the proportion of incorrect solutions in an empirical sample) is high (e.g., "What is the capital of Mongolia?" correct answer: Ulaanbaatar) or if a lot of similar questions need to be answered (e.g., questions about the capitals of all countries in the world). On the other hand, questions that require more complex cognitive elaboration processes, such as evaluating evidence (e.g., regarding the suitability of DNA analysis methods) with regard to some standards (e.g., error-proneness regarding lab results or statistical analyses), are always considered more complex – independently of task difficulty.

Research Questions – Hypotheses

In the present study, three research questions were addressed based on these theoretical considerations. The first questions pertain to adaptation to task complexity: (a) Do students *discriminate* between tasks of different complexity? We predicted that students would discriminate significantly between *Task Levels* of different complexity in their task definitions, goals and plans and that this would be evident in their answers in the COPES-questionnaire (Hypothesis 1). In short (for more information see method section), scales of the COPES questionnaire indicate task definitions, goals and plans either for deep or for superficial processing. (b) Do students *calibrate* their task definitions, goals and plans to task complexity? We predicted that students would calibrate their answers on the COPES-questionnaire significantly to task complexity and this would be evident by a systematic relationship between students' answers in the COPES-questionnaire and *Task Levels* of different complexity. More specifically, we predicted that students would judge all variables indicating deep processing more important for more complex tasks and all variables indicating superficial processing less important for more complex tasks (Hypothesis 2).

Further question pertain to the impact of personal characteristics: (c) Are these *discrimination* and *calibration* processes related to students' learner characteristics? First, we predicted effects of *epistemological beliefs*. More specifically, we predicted that students with more sophisticated epistemological beliefs would judge all variables indicating deep processing more important across all tasks and would judge all variables indicating superficial processing less important across all tasks (Hypothesis 3; main effects). Additionally, we predicted that these differences between "sophisticated" and "naïve epistemological beliefs would be more pronounced in more complex *Task Levels* (Hypothesis 4; interaction effects). Second, we predicted similar effects of prior domain-specific knowledge. Prior domain-specific knowledge showed a crucial impact on learning processes in most other studies (Lind & Sandmann, 2003; McDonald & Stevenson, 1998). The COPES model (Winne & Hadwin, 1998) predicts a similar functional relationship for prior domain-specific knowledge as for epistemological beliefs. In this study, we systematically compared two Prior Knowledge Groups: Biology students with high prior biology knowledge and humanities students with almost no prior biology knowledge. More specifically, we predicted that students with higher prior domain-specific knowledge would judge all variables indicating deep processing more important across all tasks and would judge all variables indicating superficial processing less important across all tasks (Hypothesis 5; main effects). Additionally, we predicted that the differences between high and low prior domain-specific knowledge would be more pronounced in more complex *Task Levels* (Hypothesis 6; interaction effects).

METHOD

Procedure

The present study was conducted in two sessions. During the first online session, students filled in questionnaires about their domain-general (*EBI*; Jacobson & Jehng, 1999) and domain-dependent (*CAEB*; Stahl & Bromme, 2007) epistemological beliefs, which took them about 15 minutes. Sixty-five biology students and 64 humanities students completed these online-questionnaires. The second face-to-face session was held in groups with a minimum of 3 students and a maximum group size of 12 and lasted approximately one hour. Not all students continued; 52 biology students (80% of the original sample) and 50 humanities students (78% of the original sample) participated in this second session where they had to fill in paper-pencil-questionnaires. First, students had to answer a *Short Knowledge Test* about molecular genetics and the *Self-Rated Prior Biology Knowledge* item. Then, all students read a factual introduction[1] to molecular biology which adequately contextualized students to the topic of "genetic fingerprinting". In the main part of this session, students evaluated six learning tasks of different complexity according to Bloom's revised taxonomy (Anderson et al., 2001: *remember, understand, apply, analyze, evaluate*, and *create*) with the *COPES-questionnaire*. Tasks were presented in random order.

Participants

Students who participated in both sessions constitute the final sample. All students were selectively recruited to ensure two levels of biology knowledge. Biology students were recruited during regular courses in biology; humanities students were recruited by a posting at the psychological institute. All students received 10 Euros reimbursement. Although the advanced students of biology were no "real" experts in the specific topic of "genetic fingerprinting" (Chi, 2006), they can be considered discipline experts (Rouet, Favart, Britt, & Perfetti, 1997) because they know the tools of their discipline, for example how to interpret an electrophoretogram. Students of humanities on the other hand can be considered novices (Chi, 2006).

The 52 biology students' (35 female) mean age was 22.10 years (*SD* = 2.19) and they studied on average in the 3.00rd semester (*SD* = 0.28) biology or related majors. The 50 humanities students' (43 female) mean age was 23.61 years (*SD* = 4.47) and they studied on average in the 4.24th semester (*SD* = 2.24) psychology or other humanities. Biology students significantly outperformed humanities students on a *Short Knowledge Test* (see below; t (100) = 20.63, p < .001, Cohen's d = 4.08; biology students: M = 7.23, *SD* = 1.06; humanities students: M = 2.16, *SD* = 1.40; 8 points maximum). Furthermore, they also possessed higher *Self-Rated Prior Biology Knowledge* (see below; t (99) = 5.60, p < .001, Cohen's d = 1.11; biology students: M = 2.79, *SD* = 0.73; humanities students: M = 1.90, *SD* = 0.87; on a scale from 1 (very low) to 5 (very high)). Thus, these two quasi-experimental groups of students (*Prior Knowledge Groups:* biology vs. humanities students) were used as predictor variable in all subsequent analyses to explore the effects of prior domain knowledge.

Measures

Short Knowledge Test Background knowledge in molecular biology was tested with eight multiple-choice questions (Cronbach's α = .89) that were developed with the help of a domain expert. Sample item: "What does the abbreviation PCR stand for?" Multiple-choice options: "(1) Protein Coupling Reaction, (2) Phosphate Chain Reaction, (3) *Polymerase Chain Reaction*, (4) Polysaccharide Chain Reaction, (5) Phosphate Coupling Reaction, or (6) I don't know." Each question had one correct answer (in the example in italics).

Self-Rated Prior Biology Knowledge Students were also asked to self-assess their own knowledge in genetics with the item: "I estimate my prior domain-specific knowledge in genetics to be" Answers could be given on a Likert-type scale ranging from 1 (very low) to 5 (very high).

Epistemological beliefs For the measurement of epistemological beliefs the distinction between explicit-denotative and associative-connotative aspects of epistemological beliefs was used. The distinction has been proposed by Stahl and Bromme (2007) because of the often reported problems of measuring epistemological beliefs in a reliable way (Niessen et al., 2004; Strømsø et al., 2008). The two aspects of epistemological beliefs are not necessarily in accordance with each other and they have to be measured separately.

The *explicit-denotative* aspects of epistemological beliefs were measured with an adapted version of the Epistemological Beliefs Instrument (EBI; Jacobson & Jehng, 1999) that comprises items such as "If scientists try hard enough, they can find the answer to almost every question" that had to be rated on a Likert-type scale ranging from 1 (totally disagree) to 7 (totally agree). The original instrument consists of 61 items. However, for this study only items that refer to epistemology in a strict sense were selected, more specifically from the scales *certainty of knowledge* (9 items), *omniscient authority* (5 items), and *simple view of learning* (3 items). Furthermore, we added 4 items from Wood and Kardash's (2002) questionnaire and 2 items from our own lab. The exploratory factor analysis of these 23 items of the adapted EBI applied to the sample of the present study yielded one factor explaining 35.91 % of variance; this scale was labelled *EBI-definitude.* This scale measures whether students assume that absolute answers are attainable or whether knowledge is indefinite (9 items, Cronbach's α = .76; sample items are "For most scientific research questions there is only one right answer.", "Most words have one clearly defined meaning.").

To capture the *associative-connotative* aspects of epistemological beliefs about knowledge in the domain of genetics a semantic differential, namely the Connotative Aspects of Epistemological Beliefs (CAEB; Stahl & Bromme, 2007), was used. This instrument consists of 24 pairs of antonymous adjective as items; on each item the degree of association could be rated on a 7-point scale. Sample item: "Knowledge in genetics is: simple (1) – complex (7)". The exploratory factor analysis of the 24 items of the CAEB yielded two factors explaining 50.39 % of the variance, namely CAEB-texture and CAEB-variability. The factor *CAEB-texture* encompasses beliefs about the structure and accuracy of knowledge (9 items loaded on this factor). Sample items are "Knowledge in genetics is: from 1 (precise / sorted / exact / etc.) to 7 (imprecise / unsorted / vague / etc.)"; Cronbach's α = .82. The factor *CAEB-variability* encompasses beliefs about the stability and dynamics of knowledge (5 items loaded on this factor). Sample items are "Knowledge in genetics is: from 1 (irrefutable / flexible / completed / etc.) to 7 (refutable / inflexible" / uncompleted / etc.)"; Cronbach's α = .67.

These three factors, namely *EBI-definitude, CAEB-texture*, and *CAEB-variability*, were used as predictor variables in all relevant subsequent analyses to explore the effects of epistemological beliefs.

Tasks Six tasks of different complexity according to Bloom's revised taxonomy (Anderson et al., 2001) were presented. This taxonomy distinguishes between six task classes affording cognitive processes of different complexity (in order of ascending complexity): (a) remember, (b) understand, (c) apply, (d) analyze, (e) evaluate, and (f) create. For the present study, one task for each Bloom category was constructed and selected in a cyclic process. First, two experts in biology searched through relevant textbooks for adequate tasks and constructed additional tasks for all categories.

Second, the resulting pool of about 100 tasks was independently categorized by five raters into the six Bloom categories; these raters were blind to the experts' categorization. For 39 tasks all raters immediately agreed, for further 25 tasks four of the raters agreed; the remaining tasks were either rephrased and re-categorized (15 tasks) or deleted from the pool. Third, based on content considerations six tasks per Bloom category were selected for an exploratory study (Stahl et al., 2006). Fourth, for the present study only the most prototypical task for each Bloom category was chosen based on participants' categorizations in the Stahl et al. (2006) study.

As simplest *remember* task a multiple-choice question about how to split DNA was selected; the answer only required recall of facts. As *understand* task a multiple-choice question about which errors in STR (Short Tandem Repeats) profiling could cause an erroneous match was used; to answer this question an understanding of the whole process was necessary. The *apply* task required constructing a father's DNA profile from the profiles of his wife and his biological daughters in a table; knowledge about the heredity of DNA had to be applied to this concrete problem. The *analyze* task required to detail the STR analysis process step-by-step and outline potential problems; it required participants to have a detailed mental model of the whole process. The *evaluate* task asked to evaluate the impact of DNA degradation on different methods of DNA analysis in an open answer format; it required knowledge about this topic as well as critical thinking. The most complex *create* task required describing the consequences of a law change that would allow the analysis of coding DNA regions in an open answer; this task required original and creative thinking.

All six tasks were presented in random order to each participant. Participants did not solve these tasks but had to evaluate each task with the COPES questionnaire. In this study students' adaptation to these *Task Levels* of different complexity was explored (remember, understand, apply, analyze, evaluate, and create).

The COPES Questionnaire The COPES questionnaire (Stahl et al., 2006) measures students' judgments regarding their preparatory stages of self-regulated learning, namely task definition, goal setting and planning (Winne & Hadwin, 1998) and consists of 46 items. The whole questionnaire was administered in this study for each task, but only 18 items were further analysed, namely those items where participants of an exploratory study (Stahl et al., 2006) demonstrated significant discrimination and calibration. These items cover most facets of self-regulated learning (i.e., conditions, operations, evaluations, and standards).

Two of these items required short open answers; students had to estimate the number of concepts (*estimated concepts*) and the time needed for task completion (*estimated time*). One item (*Bloom classification*) had a forced-choice format with six alternative answers that represent the *Task Levels* of different complexity according to Bloom's revised taxonomy.

The remaining fifteen items were rated on 7-point Likert-type scales, mostly ranging from very unimportant (1) to very important (7); these fifteen items were subjected to an exploratory factor analysis on the present sample that explained 62% of variance and yielded three meaningful factors: *Deep Processing* (8 items, Cronbach's α = .89; sample item: "Imagine you would have to actually solve the present task. In your opinion, how unimportant or important is it to employ the learning strategy of 'elaborating deeply'?"), dealing with *Multiple Information Sources* (5 items, Cronbach's α = .82; sample item: "… In your opinion, how unimportant or important is it to concentrate on information about 'multiple perspectives'?"), and *Superficial Processing* (2 items, Cronbach's α = .63; sample item: "… In your opinion, how unimportant or important is it to employ the learning strategy of 'memorizing'?"). These results indicated that the items were not grouped together according to the facets of the COPES model but rather according to three different approaches to learning.

These three COPES factors, namely *Deep Processing, Multiple Information Sources,* and *Superficial Processing*, as well as the three single items, namely *estimated concepts, estimated time*, and *Bloom classification*, were used as dependent variables in all subsequent analyses, each repeatedly measured six times for the six *Task Level*s representing the Bloom categories. No theoretical assumptions were made about the importance of these factors and items for the tasks of different complexity; rather the students' opinions were important.

RESULTS

Descriptives and Interrelations Regarding Learner Characteristics

The two *Prior Knowledge Groups* (biology students vs. humanities students) did not differ in their domain-related epistemological beliefs measured by the CAEB. On average students believed that knowledge in genetics is quite structured (*CAEB-texture*: $M = 3.33$, $SD = .80$; on a scale from 1 = structured – 7 = unstructured) but tentative (*CAEB-variability*: $M = 3.04$, $SD = .85$; on a scale from 1 = variable – 7 = static). However, with regard to the definitude of knowledge in general (*EBI-definitude*) the *Prior Knowledge Groups* differed significantly (F (1,100) = 10.55, $p < .01$, $d = .64$): Humanities students believed much less ($M = 2.59$, $SD = .71$) in the definitude of knowledge in general than did biology students ($M = 3.09$, $SD = .84$; on a scale from 1 = knowledge is indefinite – 7 = absolute answers are attainable).

Furthermore, the domain-general scale of the EBI (*EBI-definitude*) was not correlated significantly to any of the domain-related scales of the CAEB. However, the two domain-related scales were significantly interrelated ($r = -.52$, $p < .001$): A strong belief in structured knowledge in genetics (low value on *CAEB-texture*) was related to a strong belief in static knowledge in genetics (high value on *CAEB-variability,* the inverse relationship is due to the construction of the CAEB scales. Both endpoints point to a more 'naïve' view).

Do Students Discriminate between Tasks of Different Complexity?

We hypothesized that students should discriminate between tasks of different complexity which should be evident in their significantly different answers in the

COPES questionnaire regarding different *Task Levels* (Hypothesis 1). To test this hypothesis, we computed a MANOVA for the three COPES factors (*Deep Processing; Multiple Information Sources*; and *Superficial Processing*) with *Task Levels* as repeated-measure factor. We computed similar ANOVAs for the three remaining single items (*estimated time; estimated concepts*; and *Bloom classification*). Thus, we expected seven main effects of the repeated-measure factor *Task Level* (six univariate main effects plus one multivariate main effect).

The repeated-measure MANOVA for the three COPES factors (see Table 1 descriptives; Table 2 results) showed a multivariate main effect for the repeated-measure factor *Task Levels* which was replicated univariately on each single COPES factor (*Deep Processing; Multiple Information Sources*; and *Superficial Processing*). Exploring this question in more detail, we additionally compared the adjacent *Task Levels* statistically (after Bonferroni correction with alpha $p < .01$).

Table 1.Means and standard deviations (in brackets) for all dependent variables (rows) with regard to all Task Levels (columns)

Dependent Variable	*RE*	*UN*	*AP*	*AN*	*EV*	*CR*
Deep Processing	2.62 (1.18)	3.17 (1.05)	4.62 (1.16)	5.23 (.93)	4.86 (.99)	4.76 (.92)
Multiple I. Sources	2.28 (1.04)	3.24 (1.16)	2.98 (1.05)	3.57 (1.28)	3.84 (1.07)	5.56 (.92)
Superficial Pro.	5.52 (1.51)	3.94 (1.70)	4.05 (1.39)	4.21 (1.28)	4.10 (1.24)	2.71 (1.18)
Estimated time	7:42 (13:07)	9:49 (11:47)	44:19 (76:59)	52:39 (51:26)	38:13 (61:23)	37:11 (64:46)
Estimated concepts	2.11 (2.17)	2.51 (2.18)	4.20 (4.11)	5.85 (4.89)	5.12 (4.15)	4.99 (11.07)
Bloom classification	1.17 (.48)	2.39 (1.33)	3.28 (.91)	3.31 (1.32)	3.75 (1.34)	5.05 (1.01)

Task Levels: RE = remember, UN = understand, AP = apply, AN = analyse, EV = evaluate, and CR = create; Dependent Variables: Multiple I. Sources = Multiple Information Sources and Superficial Pro. = Superficial Processing.

For the COPES factor *Deep Processing remember* and *understand* (F (1,101) = 22.91, $p < .001$, $\eta^2_p = .18$), *understand* and *apply* (F (1,101) = 126.99, $p < .001$, $\eta^2_p = .56$), *apply* and *analyze* (F (1,101) = 26.28, $p < .001$, $\eta^2_p = .21$), and *analyze* and *evaluate* (F (1,101) = 9.47, $p < .01$, $\eta^2_p = .09$) tasks differed significantly. This means that students successfully discriminated between *Task Levels* of different complexity except for the two most complex ones (*evaluate* and *create*). Furthermore, the descriptive values (see Table 1) show that they considered *Deep Processing* less important for the most complex tasks (*evaluate* and *create*) than for the moderately complex *analyze* task. For the COPES factor *Multiple Information Sources*, the following *Task Levels* differed significantly: *remember vs. understand* (F (1,101) = 70.64, p < .001, η^2_p = .41), *apply vs. analyze* (F (1,101) = 31.01, p < .001, η^2_p= .24), and *evaluate vs. create* (F (1,101) = 194.37, p < .001, η^2_p = .66). Thus, students successfully discriminated between four *Task Levels*: the simplest *remember* task, the little more complex *understand* and *apply* tasks, the moderately complex *analyze* and *evaluate* tasks, and the most complex *create* task. For the COPES factor *Superficial Processing*, the following *Task Levels* differed significantly: *remember vs. understand* (F (1,101) = 79.87, p < .001, η^2_p = .44) and *evaluate vs. create* (F (1,101) = 105.75, p < .001, η^2_p = .51). Thus, students successfully discriminated between three broader *Task Levels*: the simplest *remember* task, a range of moderately complex tasks (*understand, apply, analyze*, and *evaluate*), and the most complex *create* task (for descriptives see Table 1).

The repeated-measure ANOVAs also indicate significant effects of the repeated-measure factor *Task Levels* for each of the three remaining single items (*estimated time; estimated concepts*; and *Bloom classification*; see Table 2). Exploring these results in more detail, we report all significant differences between adjacent *Task Levels* (for descriptives see Table 1). For the variable *estimated time*, the following *Task Levels* differed significantly: *understand vs. apply* (F (1,101) = 22.72, p < .001, η^2_p = .18). Thus, students successfully discriminated between two *Task Levels*: simple tasks (*remember* and *understand*) and complex tasks (*apply, analyze, evaluate*, and *create*). For the variable *estimated concepts*, the following *Task Levels* differed significantly: *understand vs. apply* (F (1,98) = 22.70, p < .001, η^2_p = .19) and *apply vs. analyze* (F (1,98) = 18.21, p < .001, η^2_p = .16). Thus, students successfully discriminated between three *Task Levels*: simple tasks (*remember* and *understand*), the mid-complex *apply* task, and complex tasks (*analyze, evaluate*, and *create*). For the variable *Bloom classification*, the following *Task Levels* differed significantly: *remember vs. understand* (F (1,94) = 69.94, p < .001, η^2_p = .43), *understand vs. apply* (F (1,94) = 30.43, p < .001, η^2_p= .25), and *evaluate vs. create* (F (1,94) = 64.43, p < .001, η^2_p = .41). Thus, students successfully discriminated between four Task Levels: the simplest *remember* task, one little more complex *understand* task, a range of moderately complex tasks (*apply, analyze*, and *evaluate*), and the most complex *create* task.

Do Students Calibrate their Judgments to Task Complexity?

We hypothesized that students should calibrate their judgments to task complexity which should be evident in systematic relationships between students' answers in the COPES-questionnaire and *Task Levels* of different complexity (Hypothesis 2). To test this hypothesis, we computed intra-individual Goodman-Kruskal Gamma correlations *(G)* between the *Task Levels* and each dependent variable (in all cases: n = 6, for six *Task Levels*) to diagnose calibration. These correlations were

Table 2.Repeated-measure (M)ANOVAs regarding the effects of Task Levels

	F	*df*	*df error*	*p*	*partial* η^2
repeated-measure MANOVA across Task Levels for the three COPES factors					
Task Levels (multivariate) $^+$	60.22	15	87	< .001	.91
Deep Processing	148.38	5	505	< .001	.60
Multiple Information Sources	169.86	5	505	< .001	.63
Superficial Processing	62.78	5	505	< .001	.38
separate repeated-measure ANOVAs across Task Levels for the three single items					
Estimated time	17.15	5	97	< .001	.47
Estimated concepts	19.66	5	94	< .001	.51
Bloom classification	258.06	5	90	< .001	.94

$^+$Multivariate effects; all values according to Pillai's trace; univariate effects indented.

Table 3. Calibration indices indicating the relationship between the dependent variables (rows) and Task Levels of different complexity

Dependent Variable	*Index M (SD)*	*Significance*	*G*
Deep Processing	.60 (.41)	t (101) = 14.69***, d = 1.46	.54
Multiple Information Sources	.95 (.70)	t (101) = 13.70***, d = 1.36	.74
Superficial Processing	-.61 (.88)	t (101) = -7.03***, d = .69	-.55
estimated time	.65 (.55)	t (101) = 11.96***, d= 1.18	.57
estimated concepts	.51 (.72)	t (101) = 7.13***, d = .71	.47
Bloom classification	1.23 (1.02)	t (101) = 12.23,***, d = 1.21	.85

**** p < .001; M = mean; SD = standard deviation; d = Cohen's d; G = Goodman-Kruskal Gamma correlation, in this column the G values that correspond to the calibration indices ("Index") are reported (reverse Z-transformation of the mean calibration indices).*

subsequently Z-transformed into calibration indices. We determined significance by statistically testing the magnitude of these average indices against zero. We expected six calibration indices of significant size, one for each of the six dependent variables (*Deep Processing; Multiple Information Sources; Superficial Processing; estimated time; estimated concepts;* and *Bloom classification*).

We found significant calibration indices for all dependent variables (see Table 3). For example, the positive correlation of *G* = .54 between students' answers regarding *Deep Processing* and *Task Levels* indicates that students judged *Deep Processing* to be quite unimportant for simple tasks and of ascending importance for more complex tasks (for descriptives see Table 1). Similar positive relationships were detected for *Multiple Information Sources, estimated time, estimated concepts,* and *Bloom* *classification* (Table 3). The negative correlation of G = -.55 (see Table 3) between students' answers regarding *Superficial Processing and Task Levels* on the other hand indicates the following: Students judged *superficial processing* to be quite important for simple tasks and of descending importance for more complex tasks.

In addition to relative calibration (see above), *absolute calibration* was explored for students' *Bloom classifications*. This was the only instance where absolute calibration could be analysed within this study. For all other dependent variables we had no comparative standard indicating what constitutes correct answers. But for *Bloom classifications*, students' classifications could be directly compared to the correct classifications (see methods section). Students on average classified 47.17 % of the six tasks correctly (M = 2.83, SD = 1.26) which is significantly more than could be randomly expected (namely one out of six; t (97) = 14.35, p < .001, Cohen's d = 1.45). The corresponding calibration graph (see Figure 1) shows that students slightly overestimated the complexity of simpler tasks (*remember – apply*) while they underestimated the complexity of more complex tasks (*analyze - create*) compared with hypothetically perfect classifications (indicated by the "line of perfect calibration" in Figure 1).

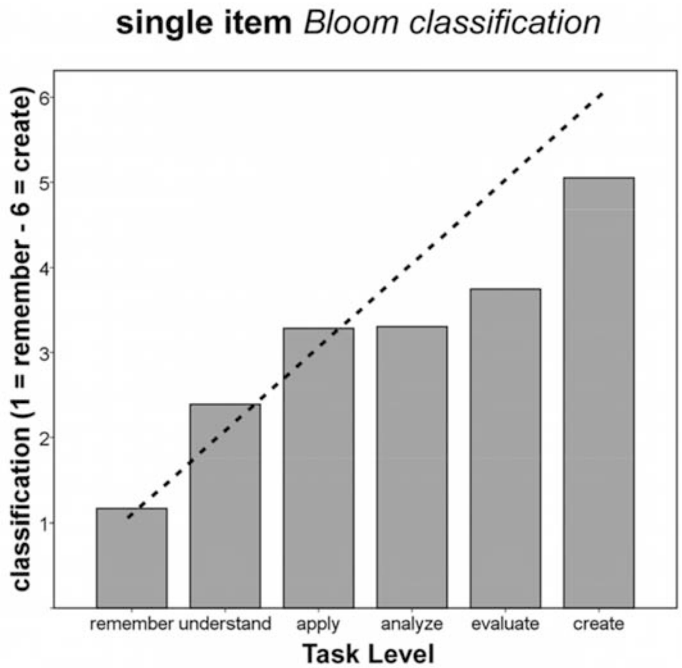

Figure 1.Calibration graph depicting students' Bloom classifications (Y-axis) as a function of Task Levels of different complexity (X-axis). The dotted line represents the hypothetical "line of perfect calibration" (perfectly correct classifications).

Are these Metacognitive Discrimination and Calibration Processes Related to Students' Learner Characteristics?

To test our hypotheses regarding learner characteristics and *discrimination*, repeated-measure analyses were computed including all learner characteristics simultaneously. More specifically, a repeated-measure MANCOVA was computed across the three COPES scales (*Deep Processing, Multiple Information Sources*, and *Superficial Processing*) with *Task Levels* as repeated-measure factor, the epistemological beliefs scales (*EBI-definitude, CAEB-variability*, and *CAEB-texture*) as covariates and the *Prior Knowledge Groups* (biology students vs. humanities students) as factor. Additionally, repeated-measure ANCOVAs were computed separately for each of the remaining single items (*estimated concepts, estimated time*, and *Bloom classification*) with the same covariates, repeated-measure and between-subject factors.

To test our hypotheses regarding learner characteristics and calibration (also see interaction effects above), correlations were computed between the calibration indices of all dependent variables (*Deep Processing, Multiple Information Sources, Superficial Processing, estimated concepts, estimated time*, and *Bloom classification*) and the epistemological beliefs scales (*EBI-definitude, CAEB-variability*, and *CAEB-texture*). Additionally, the calibration indices of all dependent variables were statistically compared between *Prior Knowledge Groups* (biology students vs. humanities students).

Note that even though in all cases all learner characteristics (epistemological beliefs scales *and* Prior Knowledge Groups) were simultaneously included in the analyses, we report the results separately: We will first report all results regarding epistemological beliefs, namely the results regarding main effects (Hypothesis 3) and the results regarding interaction effects (Hypothesis 4). Subsequently, we will report all results regarding prior domain-specific knowledge, namely the results regarding main effects (Hypothesis 5) and the results regarding interaction effects (Hypothesis 6). We will only report the significant results but we will point out the number of non-significant effects in each analysis.

Effects of Epistemological Beliefs We hypothesized that more sophisticated beliefs should be associated with judging all variables indicating deep processing more important across all tasks and with judging all variables indicating superficial processing less important (Hypothesis 3). Therefore, we expected a total of twenty-one main effects (18 univariate main effects of three epistemological belief scales regarding six dependent variables and 3 multivariate main effects of three epistemological beliefs scales).

In the MANCOVA across the three COPES factors we found a significant multivariate main effect of *CAEB-variability* (F (3,95) = 2.85, $p < .05$, $\eta^2_p = .083$) that was univariately replicated significantly on the COPES factors *Deep Processing* (F (1,97) = 5.23, $p < .05$, $\eta^2_p = .051$, Figure 2, top) and *Multiple Information Sources*

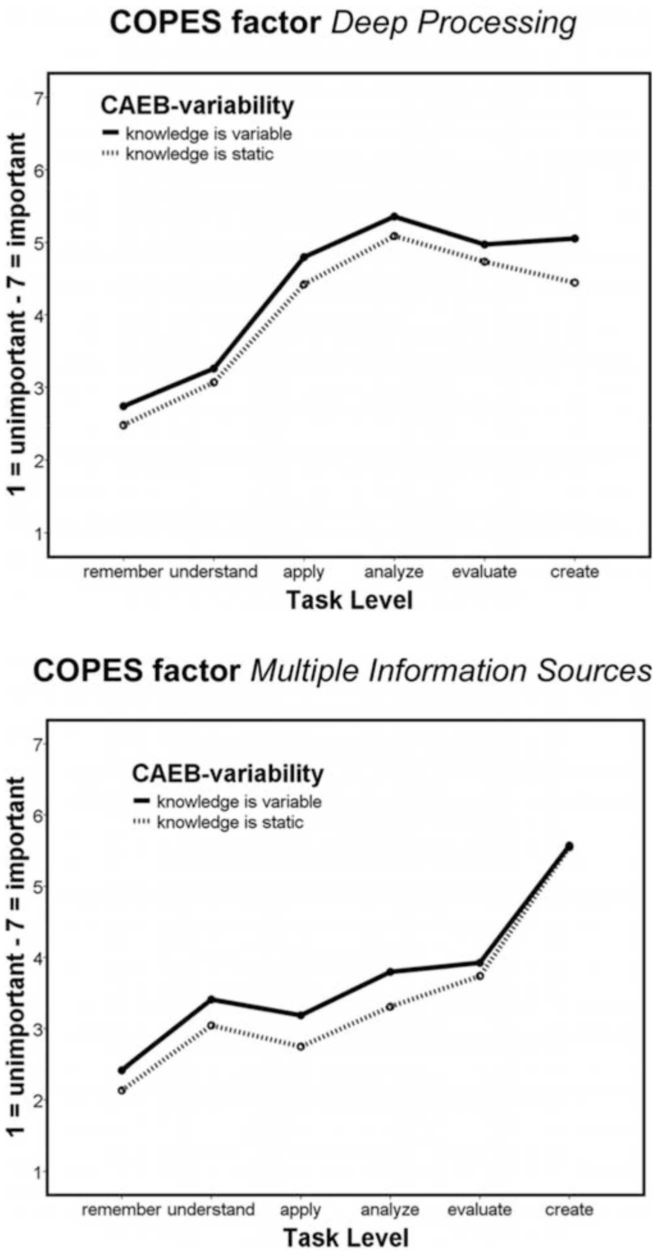

Figure 2. Calibration graphs depicting students' judgments on the COPES factors Deep Processing (top) and dealing with Multiple Information Sources (bottom) as a function of Task Levels (X-axis) and CAEB-variability (median-split; lines).

(F (1,97) = 7.58, $p < .01$, $\eta^2_p = .072$, Figure 2, bottom): Students, who considered knowledge in genetics variable (sophisticated view on *CAEB-variability*) also considered *Deep Processing* and *Multiple Information Sources* more important across all *Task Levels* than more naïve students. Note that these effects were visualized by median-splitting the scale *CAEB-variability* (Figure 2), but that *CAEB-variability* was included as covariate in the analyses!

Additionally, the ANCOVA for the single item *Bloom classification* indicates a significant main effect of *CAEB-variability* (F (1,90) = 4.59, $p < .05$, $\eta^2_p = .049$, without Figure): More sophisticated students who believed in variable knowledge in genetics classified tasks in more complex *Task Levels* (especially *analyze* tasks). To summarize: We found four significant main effects in the expected direction, all of the scale *CAEB-variability*; all other main effects of epistemological beliefs were not significant.

We hypothesized that the effects of epistemological beliefs would be more pronounced on more complex *Task Levels* (Hypothesis 4). Therefore, we expected a total of twenty-one interaction effects (18 univariate interactions between *Task Levels* and three epistemological belief scales regarding six dependent variables and 3 multivariate interactions between *Task Levels* and three epistemological beliefs scales). Additionally, we expected a total of eighteen significant correlations with calibration indices (for each of three epistemological beliefs scales with six dependent variables).

In the MANCOVA across the three COPES factors we found a significant univariate interaction between *CAEB-variability* and the repeated-measure factor *Task Levels* for the COPES factor *Multiple Information Sources* (F (5,485) = 2.34, $p < .05$, $\eta^2_p = .024$, Figure 2, bottom). The above-mentioned main effect of *CAEB-variability* was most pronounced for the *Task Levels remember* through *analyze*, while it disappeared for the more complex tasks *evaluate* and *create*. Furthermore, we found one significant correlation with a calibration index: More naïve beliefs in the definitude of knowledge in general (*EBI-definitude*) were significantly associated with higher calibration indices regarding *estimated concepts* ($r = .26$, $p = .009$). To summarize: We found two effects indicating interactions between epistemological beliefs and task complexity (*Task Levels*), both counterintuitive. All other interaction effects and effects on calibration were not significant.

Effects of Prior Domain-Specific Knowledge We hypothesized that more domain-specific knowledge should be associated with judging all variables indicating deep processing more important across all tasks and with judging all variables indicating superficial processing less important (Hypothesis 4). Therefore, we expected a total of seven main effects (6 univariate main effects of Prior Knowledge Groups regarding six dependent variables and 1 multivariate main effect of Prior Knowledge Groups). However, we found no significant main effects of prior domain-specific knowledge at all.

We hypothesized that the effects of prior domain-specific knowledge would be more pronounced on more complex *Task Levels* (Hypothesis 6). Therefore, we expected a total of seven interaction effects (6 univariate interactions between *Task Levels* and *Prior Knowledge Groups* regarding six dependent variables and 1 multivariate interaction between *Task Levels* and *Prior Knowledge Groups*). We found a significant multivariate interaction between the *Task Levels* and *Prior Knowledge Groups* (*F* (15,83) = 2.03, *p* < .05, η^2_p = .268) that was univariately only replicated on the COPES factor *Deep Processing* (*F* (5,485) = 2.94, *p* < .05, η^2_p = .029, Figure 3): Biology students judged *Deep Processing* to be of ascending importance from *remember* tasks through *analyze* tasks and their judgments reached a plateau for *analyze, evaluate* and *create* tasks. Humanities students did not discriminate on such a fine-grained level. They judged *Deep Processing* to be quite unimportant for *remember* and *understand* tasks and quite important for all more complex tasks. Furthermore, we found one significant difference in calibration indices (*t* (100) = 2.09, *p* = .039, *d* = .41): Biology students (calibration: *M* = .65, *SD* = .83) displayed significantly higher calibration indices with regard to *estimated concepts* than humanities students (calibration: *M* = .36, *SD* = .56). To summarize: We found

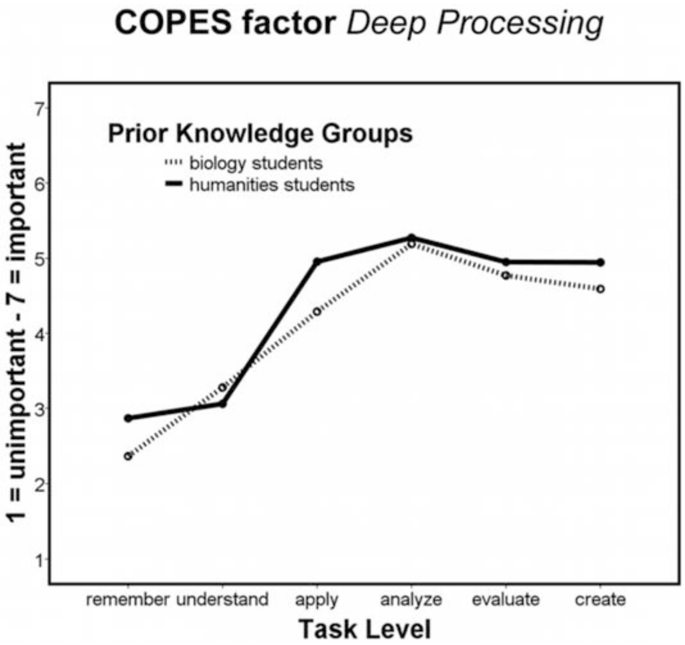

Figure 3. Calibration graph depicting students' judgments on the COPES factor Deep Processing as a function of Task Levels (X-axis) and Prior Domain Knowledge Groups (biology students vs. humanities students; lines).

two effects indicating interactions between prior domain-specific knowledge and task complexity. All other interaction effects and effects on calibration were not significant.

DISCUSSION

Discrimination and Calibration

The empirical data of the present study confirm Hypothesis 1. The repeated-measure factor *Task Levels* elicited significant main effects on *all* dependent variables (*Deep Processing, Multiple Information Sources, Superficial Processing, estimated concepts, estimated time*, and *Bloom classification*). This means that students in fact discriminate between tasks of different complexity as evident in their significantly different answers on the COPES questionnaire.

The empirical data of the present study also confirm Hypothesis 2. *Task Levels* of different complexity were significantly correlated with scores on *all* dependent variables. This means that students in fact calibrate their answers in the COPES questionnaire systematically to task complexity. More specifically, they consider all indicators of deep processing (*Deep Processing, Multiple Information Sources, estimated concepts, estimated time*, and *Bloom classification*) more important for more complex tasks and they consider indicators of superficial processing (*Superficial Processing*) less important for more complex tasks.

Therefore, the results regarding the first two research questions are consistent with the COPES-model (Winne & Hadwin, 1998) that assumes that students systematically adapt their learning process to external conditions. Furthermore, these results are mostly consistent with those of previous empirical studies about task complexity (e.g. Gall, 2006; Klayman, 1985; Rouet, 2003; Winne & Jamieson-Noel, 2003). Most of these studies focused on the *enactment* of learning strategies and indicate that learners demonstrate good self-regulation for simple tasks but less adequate self-regulation for complex tasks. The results of this study are consistent because in all cases learners processed different complex tasks differently and systematically adapted their (planned) behavior to task complexity. However, the results of this study are inconsistent with regard to the quality of students' self-regulation: While results from other studies indicate insufficient self-regulation for complex tasks the results of this study indicate that students are well aware of the special demands of complex tasks and plan to use adequate approaches. One potential explanation for this inconsistency concerns the different stages of learning: Students might be able to *plan* adequate self-regulation based on their adequate metacognitive knowledge about tasks and strategies (this study) but they might be unable to *enact* the planned approaches, for example due to cognitive overload or due to production or motivation deficits (other studies).

One directly related open issue concerns the *absolute* quality of students' calibration. Even though students in general are quite successful at discriminating and calibrating (see above) they might be still far from perfect. Overestimating the complexity of simple tasks might not be detrimental for learning, just not be the most parsimonious way to solve these simple task. Misjudging the complexity of more complex tasks on the other hand might have more detrimental effects. Not only would the answer be less adequate, but also the gained understanding would be more superficial than required. Data from this study (as well as from the corresponding exploratory study; Stahl et al., 2006) tentatively indicates that students might in fact underestimate the complexity of very complex tasks – which would be in line with the finding of less adequate self-regulation for more complex tasks in other studies (see above). For example, the calibration graph depicting students' *absolute calibration* for the item *Bloom classification* indicates that students classify complex tasks into less complex *Task Levels* than warranted (Figure 1). However, this interpretation requires further caution because of our definition of task complexity: Bloom's revised taxonomy (Anderson et al., 2001) assumes a cumulative hierarchy. But empirical results – testing Bloom's original taxonomy (Bloom et al., 1956) – show that the most complex tasks can often not be discriminated with regard to complexity or difficulty (Kreitzer & Madaus, 1994; Kunen, Cohen, & Solman, 1981). On the other hand empirical results strongly support the hierarchical order of less complex tasks, especially for understand, apply, analyze, and create (Gierl, 1997; Kreitzer & Madaus, 1994; Kunen et al., 1981). To conclude this argument: Bloom's revised taxonomy may define task complexity a bit too fine-grained because for some complex levels very similar cognitive processes might be adequate for students. However, this potential problem does not invalidate our conclusion that students in general could – or probably should – consider indicators of deep processing even more important for complex tasks than they currently do – for *analyze* through *create* tasks.

Another issue concerns methodology transfer of calibration measures: Recall the major conceptual differences between traditional conceptualizations of calibration (i.e., accuracy of metacognitive judgments regarding one's own performance) and our conceptualization of calibration (i.e., alignment between students' task definitions, goals and plans and the external variable task complexity). Presumably, learners possess more metacognitive awareness about their own internal cognitive processes (traditional conceptualizations) than about the fit of these processes with the external world (our conceptualizations). Therefore, if we compared these conceptually different indices we would expect higher indices of relative calibration in traditional calibration research. Thus, it is surprising that we detected indices of relative calibration (Goodman-Kruskal Gamma correlations) within this new application context that range from $G = .47$ to $G = .85$. The size of these calibration indices would even be considered substantial if compared to calibration indices from the traditional calibration paradigm (e.g. $G =.38$ for immediate and $G = .90$ for delayed confidence judgments; Nelson & Dunlosky, 1991).

Epistemological Beliefs

The empirical data of the present study partly confirm Hypothesis 3 (main effects of epistemological beliefs). More sophisticated beliefs in variable knowledge in genetics (*CAEB-variability*) were significantly associated with judging variables indicating deep processing more important (for the dependent variables *Deep Processing, Multiple Information Sources*, and *Bloom classification*). However, no significant effects were detected for other epistemological beliefs scales (*CAEB-texture* and *EBI-definitude*) or other dependent variables (*Superficial Processing, estimated time*, and *estimated concepts*). To summarize: All detected main effects of epistemological beliefs (n = 4) point in the hypothesized direction, but the majority of the hypothesized effects was not significant (n = 17).

Regarding the significant effects, most likely students who believe that knowledge is variable and dynamic automatically consider all kinds of tasks more complex per se (effect on *Bloom classification*). In order to counteract this perceived complexity and in order to adequately deal with the perceived variability of knowledge they might plan deep elaboration approaches (effects on *Deep Processing* and *Multiple Information Sources*). It could be said that these sophisticated students discriminated between tasks on a higher level. These results are in line with other empirical results indicating beneficial main effects of sophisticated beliefs (Bartholomé et al., 2006; Kardash & Scholes, 1996; Mason & Boscolo, 2004; Mason & Scirica, 2006; Schommer, 1990; Schommer, Crouse, & Rhodes, 1992; Schommer-Aikins & Hutter, 2002; Muis, 2007). For example, in other studies concentrating on the *preparatory* stages of learning, students with sophisticated beliefs perceived the affordances of ill-structured tasks more accurately (King & Kitchener, 2002) and set more adequate goals (Bråten & Strømsø, 2004; Ryan, 1984).

However, the number of non-significant main effects, especially regarding other dimensions of epistemological beliefs is surprising. There were no significant (main) effects of connotative beliefs about the structure of knowledge in genetics (*CAEB-texture*) and also of the denotative beliefs about the definitude of knowledge in general (*EBI-definitude*). Possibly beliefs about structural aspects (*CAEB-texture*) of knowledge in genetics have been conceived by our subjects as issues which apply to the field of genetics in general, while issues of variability (*CAEB-variability*) might be more topic-specific and therefore they might have been more important in order to decide how different learning tasks should be tackled differently. Note that in this study epistemological beliefs were also measured in a rather abstract way, especially regarding *EBI-definitude* which was measured for knowledge in general. This also might explain why there were weaker effects than we would have expected based on our predictions. Note, that these explanations may also be relevant for explaining the non-significant interactions between epistemological beliefs and task complexity (see below).

The empirical data of the present study do not confirm Hypothesis 4 (interaction between epistemological beliefs and task complexity). We expected that the effects of epistemological beliefs would be more pronounced regarding more complex *Task Levels*. However, we found two effects that explicitly contradicted this expectation, namely the effects of *CAEB-variability* on *Multiple Information Sources* disappeared for the most complex *Task Levels* and more "naïve" beliefs in definite knowledge (*EBI-definitude*) were associated with higher calibration indices regarding *estimated concepts*. To summarize: All detected interaction or calibration effects of epistemological beliefs (*n* = 2) point in directions contrary to our hypotheses; the majority of hypothesized effects was not significant (*n* = 20 interaction effects; *n* = 17 correlations with calibration indices).

These effects are inconsistent with our predictions as well as with previous research findings. In this study, students with more naïve epistemological beliefs appear to be better at adapting their task definitions, goals and plans to task complexity while sophisticated students showed less flexibility. On the other hand, we assumed theoretically that students with more sophisticated beliefs should be more flexible in their adaptations to task complexity (Hammer & Elby, 2002). Consistent with this theoretical assumption, previous empirical studies investigating the relationship between students' epistemological beliefs and their calibration, found that sophisticated beliefs in gradual learning (*quick learning*, Schommer, 1990) as well as in complex knowledge (*simple knowledge*, Schommer et al., 1992) were associated with less overestimation of comprehension. Furthermore, the corresponding exploratory study from our lab (Stahl et al., 2006) also demonstrated that sophisticated beliefs were associated with better *calibration* indices in the preparatory stages of learning.

One potential explanation for the counterintuitive effects detected in this study is related to the measurement of epistemological beliefs: The scale *EBI-definitude* reaches from views that knowledge is definite (naïve *absolutist*) to views that knowledge is indefinite (sophisticated *relativist*) but does not capture most sophisticated flexible *evaluativist* epistemologies (Kuhn, Cheney, & Weinstock, 2000). Such an evaluativist position with regard to *EBI-definitude* would mean that although knowledge in general is considered indefinite such a person would be aware that some pieces of knowledge are well-validated by scientific inquiry and thus almost absolute answers are attainable in some cases. Most likely, students with such epistemological beliefs would give judgments in the mid-range of the scale *EBI-definitude*. The frequency distribution for *EBI-definitude* reveals that judgments in this sample range from very indefinite conceptualizations (*relativist*) to moderately definite ones (*probably evaluativist*); no very definite judgments were given. Students with moderately definite views on *EBI-definitude* – probably the most sophisticated students according to this proposed explanation – possess higher *calibration* indices than students who considered knowledge very indefinite.

Regarding epistemological beliefs, we conclude that epistemological beliefs elicited fewer effects than predicted, but that our predictions of main effects were correct, at least regarding *CAEB-variability*: Sophisticated beliefs in variable knowledge in genetics were mainly associated with judging indicators of deep processing more important across all tasks. These effects are consistent with our theoretical assumption that epistemological beliefs foster learning because they entail general assumptions about the forthcoming knowledge and task structures which have to be dealt with by the learner. Of course, the reported relationships between epistemological beliefs and task definitions, goals and plans are only correlational. Therefore we conceive the results with some reserve as evidence for our theoretical proposition about epistemological beliefs as standards for the calibration in the preparatory phases of learning as proposed in the COPES model.

Prior Domain-Specific Knowledge

The empirical data of the present study does not confirm Hypothesis 5 (main effects of prior domain-specific knowledge). We expected that biology students with high prior domain-specific knowledge would judge indicators of deep processing more important and indicators of superficial processing less important across all tasks. However, we found none of the six expected main effect of *Prior Knowledge Groups*. The empirical data of the present study show an unexpected pattern regarding Hypothesis 6 (interaction between prior domain-specific knowledge and task complexity). We expected that the effects of prior knowledge would be more pronounced on more complex *Task Levels*. However, we found two effects just indicating more fine-grained and differentiated calibration of biology students. *Prior Knowledge Groups* showed an interaction with *Task Levels* on the COPES factor *Deep Processing* indicating more fine-grained discrimination of biology students. Furthermore, biology students displayed higher calibration indices with regard to *estimated concepts*. To summarize: The detected interaction or calibration effects of prior domain-specific knowledge show more fine-grained discrimination for students with higher domain-specific knowledge – which differs from the predicted pattern of interaction; however the majority of hypothesized effects was not significant.

Prior knowledge might have helped students to perceive more fine-grained nuances of differences in tasks while students without adequate domain-specific knowledge might have based their judgments on surface cues. These results are mostly consistent with those of previous empirical studies demonstrating that prior domain-specific knowledge has little quantitative impact (consistent with the small number of detected effects) but some qualitative impact (consistent with the detected effects) on *planning* processes: Experts seem to use more elaborate criteria to evaluate tasks and seem to judge task difficulty more accurately (Chi, 2006; Lodewyk & Winne, 2005). However, considering the ubiquitous impact of prior domain-specific knowledge on learning processes detected in other empirical studies, prior domain-specific knowledge had surprisingly little impact on students' preparatory stages of self-regulated learning in this study. A potential explanation concerns the domain-specificity versus domain-generality of expertise: Students' *task definitions, goals and plans* might be more dependent on domain-general approaches to learning (e.g. students' metacognitive knowledge about tasks and adequate strategies) than on prior domain-specific knowledge. However, we assume that prior domain-specific knowledge might become more relevant in subsequent stages of learning.

IMPLICATIONS

These results imply that students are able to successfully monitor tasks with regard to complexity and seem to know reasonably well what kind of *task definitions, goals and plans* are adequate. Of course, it cannot be taken for granted that the planning and anticipation processes which were scrutinized here do really result in appropriate learning behaviour. Thus, if students should fail to enact appropriate strategies in the subsequent stages of learning, this should not be attributed to monitoring or knowledge deficits, but rather to production or motivation deficits.

If these findings could be corroborated in further studies it would have some practical implications. If students (at least of this age group) are able to apprehend the complexity of tasks in advance, such capabilities could be used in instruction. In order to make students aware about their pre-existing knowledge and ideas (sometimes also: about their misconceptions) it might be helpful to ask them for reflections about the next tasks, similarly to the procedure with the COPES questionnaire of this study. Asking students why they judge some tasks as less complex than others and asking them what they think about the knowledge laying before them, might be a successful teaching approach just because it can build on the calibration capabilities which became evident in this study. Furthermore they could be asked about their ideas with regard to the nature of the knowledge which they have got to acquire next. While our findings with regard to the relationship between such beliefs and calibration were mixed, they nevertheless allow for the conclusion that thinking about the forthcoming learning tasks involves some epistemological belief aspects. Again such relationships could be made more aware by explicit discussing denotative as well as connotative aspects of students' ideas about the knowledge they have to acquire next. We have shown that even such general associations about the variability of knowledge as they were measured here (with the CAEB) are related to the choice of study strategies. Therefore it should be feasible to use these associations as a topic of instructional discussions.

NOTES

[1] Two versions of this introduction were administered, but because this experimental treatment elicited no significant effects, we ignored this factor subsequently. More specifically, we matched two groups of participants with regard to their prior biology knowledge and their epistemological beliefs based on the results obtained in the first online session. One matched sub-sample read a neutral version of the introduction and the other sub-sample an epistemological version that was enriched with comments about epistemological issues and which was intended to elicit more sophisticated beliefs. As a treatment check the CAEB was re-administered after this epistemological sensitization. However, we found no significant differences in epistemological beliefs after this treatment. Therefore, we ignored this attempted experimental manipulation in all subsequent analyses.

REFERENCE LIST

Anderson, L. W., Krathwohl, D. R., Airasian, P. W., Cruikshank, K. A., Mayer, R. E., Pintrich, P. R., et al. (2001). *A taxonomy for learning, teaching, and assessing. A revision of Bloom's taxonomy of educational objectives.* New York: Longman.

Bartholomé, T., Stahl, E., Pieschl, S., & Bromme, R. (2006).What matters in help-seeking? A study of help effectiveness and learner-related factors. *Computers in Human Behavior, 22*, 113-129.

Bendixen, L. D., & Hartley, K. (2003). Successful learning with hypermedia: The role of epistemological beliefs and metacognitive awareness. *Journal of Educational Computing Research, 28*(1), 15-30.

Bloom, B. S., Engelhart, M. D., Furst, E. J., Hill, W. H., & Krathwohl, D. R. (1956). *Taxonomy of educational objectives: Handbook I: Cognitive domain.* New York: David McKay.

Bråten, I., & Strømsø, H. I. (2004). Epistemological beliefs and implicit theories of intelligence as predictors of achievement goals. *Contemporary Educational Psychology, 29*, 371-388.

Bråten, I., Strømsø, H. I., & Samuelstuen, M. S. (2005). The relationship between internet-specific epistemological beliefs within internet technologies. *Journal of Educational Computing Research, 33*(2), 141-171.

Bromme, R., Kienhues, D., & Porsch, T. (2009). Who knows what and who can we believe? Epistemological beliefs are beliefs about knowledge (mostly) attained from others. In L. D. Bendixen, & F. C. Feucht (Eds.), *Personal Epistemology in the Classroom: Theory, Research, and Implications for Practice* (pp. 163-193). Cambridge: Cambridge University Press.

Bromme, R., Kienhues, D., & Stahl, E. (2008).Knowledge and epistemological beliefs: An intimate but complicate relationship. In M. S. Khine (Ed.), *Knowing, Knowledge, and Beliefs: Epistemological Studies Across Diverse Cultures* (p.423 - 444). New York: Springer.

Bromme, R., Pieschl, S., & Stahl, E. (2010). Epistemological beliefs are standards for adaptive learning: A functional theory about epistemological beliefs and metacognition. *Metacognition and Learning*, 5(1), 7-26.

Buehl, M. (2008). Assessing the multidimensionality of students' epistemic beliefs across diverse cultures. In M. S. Khine (Ed.), *Knowing, knowledge and beliefs. Epistemological studies across diverse cultures* (pp. 65-112). New York: Springer.

Buehl, M. M., & Alexander, P. A. (2001). Beliefs about academic knowledge. *Educational Psychology Review, 13*(4), 385-418.

Cano, F. (2005). Epistemological beliefs and approaches to learning: Their change through secondary school and their influence on academic performance. *British Journal of Educational Psychology*, 75, 203-221.

Chi, M. T. H. (2006). Two approaches to the study of experts' characteristics. In K. Ericsson, N. Charness, P. J. Feltovich, & R. R. Hoffman (Eds.), *Cambridge handbook of expertise and expert performance* (pp. 21-30). New York: Cambridge University Press.

Dahl, T. I., Bals, M., & Turi, A. L. (2005). Are students' beliefs about knowledge and learning associated with their reported use of learning strategies? *British Journal of Educational Psychology, 75*, 257-273.

Elby, A., & Hammer, D. (2001). On the substance of a sophisticated epistemology. *Science Education, 85*(5), 554-567.

Gall, J. (2006). Orienting tasks and their impact on learning and attitudes in the use of hypertext. *Journal of Educational Multimedia and Hypermedia, 15*(1), 5-29.

Gierl, M. J. (1997). Comparing cognitive representations of test developers and students on a mathematics test with Bloom's taxonomy. *The Journal of Educational Research, 91*(1), 26-32.

Greene, J. A., & Azevedo, R. (2007). A theoretical review of Winne and Hadwin's model of self-regulated learning: New perspectives and directions. *Review of Educational Research, 77*(3), 334-372

Greene, J. A., Azevedeo, R., & Torney-Purta, J. (2008). Modeling epistemic and ontological cognition: Philosophical perspectives and methodological directions. *Educational Psychologist, 45*(3), 142-160.

Hammer, D., & Elby, A. (2002). On the form of a personal epistemology. In B. K. Hofer, & P. R. Pintrich (Eds.), *Personal epistemology. The psychology of beliefs about knowledge and knowing* (pp. 169-190). Mahwah, NJ: Lawrence Erlbaum Associates.

Hofer, B. K. (2004). Epistemological understanding as a metacognitive process: Thinking aloud during online-searching. *Educational Psychologist, 39*(1), 43-55.

Hofer, B. K., & Pintrich, P. R. (1997). The development of epistemological theories: Beliefs about knowledge and knowing and their relation to learning. *Review of Educational Research, 67*(1), 88-140.

Jacobson, M. J., & Jehng, J.-C. (1999). *Epistemological beliefs instrument: Scales and items*. Retrieved June 20, 2007, from http://mjjacobson.net/publications/Epist_Beliefs_ Instrument98.PDF.

Jacobson, M. J., & Spiro, R. J. (1995). Hypertext learning environments, cognitive flexibility, and the transfer of complex knowledge: An empirical investigation. *Journal of Educational Computing Research, 12*(4), 301-333.

Kardash, C. M., & Howell, K. L. (2000). Effects of epistemological beliefs and topic-specific beliefs on undergraduates' cognitive and strategic processing of dual-positional text. *Journal of Educational Psychology, 92*(3), 524-535.

Kardash, C. M., & Scholes, R. J. (1996). Effects of preexisting beliefs, epistemological beliefs, and need for cognition on interpretation of controversial issues. *Journal of Educational Psychology, 88*(2), 260-271.

King, P. M., & Kitchener, K. S. (2002). The reflective judgment model: Twenty years of research on epistemic cognition. In B. K. Hofer, & P. R. Pintrich (Eds.), *Personal epistemology. The psychology of beliefs about knowledge and knowing* (pp. 37-62). Mahwah, NJ: Lawrence Erlbaum Associates.

Kitchener, K. S. (1983). Cognition, metacognition, and epistemic cognition: A three-level model of cognitive processing. *Human Development, 26*, 106-116.

Klayman, J. (1985). Children's decision strategies and their adaptation to task characteristics. *Organizational Behavior and Human Decision Processes, 35*, 179-201.

Kreitzer, A. E., & Madaus, G. F. (1994). Empirical investigations of the hierarchical structure of the taxonomy. In L. W. Anderson, & L. A. Sosniak (Eds.), *Bloom's taxonomy: A forty-year retrospective* (pp. 64 - 82). Chicago: University of Chicago Press.

Kuhn, D. (2000). Metacognitive development. *Current Directions in Psychological Sciences, 9*(5), 178-181.

Kuhn, D., Cheney, R., & Weinstock, M. (2000). The development of epistemological understanding. *Cognitive Development, 15*(3), 309-328.

Kunen, S., Cohen, R., & Solman, R. (1981). A levels-of-processing analysis of bloom's taxonomy. *Journal of Educational Psychology, 73*(2), 202-211.

Lind, G., & Sandmann, A. (2003). Lernstrategien und Domänenwissen [learning strategies and domain knowledge]. *Zeitschrift für Psychologie, 211*(4), 171-192.

Lodewyk, K. R., & Winne, P. H. (2005). Relations among the structure of learning tasks, achievement, and changes in self-efficacy in secondary students. *Journal of Educational Psychology, 97*(1), 3-12.

Mason, L., & Boscolo, P. (2004). Role of epistemological understanding and interest in interpreting a controversy and in topic-specific belief change. *Contemporary Educational Psychology, 29*, 103-128.

Mason, L., & Scirica, F. (2006). Prediction of students' argumentation skills about controversial topics by epistemological understanding. *Learning and Instruction, 16*(5), 492-509.

McDonald, S., & Stevenson, R. J. (1998). Navigation in hyperspace: An evaluation of the effects of navigational tools and subject matter expertise on browsing and information retrieval in hypertext. *Interacting with Computers, 10*, 129-142.

Muis, K. R. (2007). The role of epistemic beliefs in self-regulated learning. *Educational Psychologist*, *42*(3), 173-190.

Neber, H., & Schommer-Aikins, M. (2002). Self-regulated science learning with highly gifted students: The role of cognitive, motivational, epistemological, and environmental variables. *High Ability Studies, 13*(1), 59-74.

Nelson, T. O. (1996). Gamma is a measure of the accuracy of predicting performance on one item relative to another item, not of the absolute performance on an individual item. Comments on Schraw (1995). *Applied Cognitive Psychology, 10*, 257-260.

Nelson, T. O., & Dunlosky, J. (1991). When people's judgements of learning (JOLs) are extremely accurate at predicting subsequent recall: the "delayed-JOL effect". *Psychological Science, 2*(4), 267-270.

Niessen, T.,Vermunt, J. D., Abma, T., Widdershoven, G., & van der Vleuten, C. (2004). On the nature and form of epistemologies: Revealing hidden assumptions through an analysis of instrument design. *European Journal of School Psychology, 2*(1-2), 39-64.

Pieschl, S. (2009). *Metacognitive calibration – An extended conceptualization and potential applications. Metacognition and Learning*, *4(1), 3-31*.

Pieschl, S., Stahl, E., & Bromme, R. (2013). Adaptation to context as core component of self-regulated learning. The example of complexity and epistemic beliefs. In R. Azevedo & V. Aleven (Eds.), *International Handbook of Metacognition and Learning Technologies*, Springer International Handbooks of Education 26 (pp. 53-65). New York: Springer.

Rouet, J.-F. (2003). What was I looking for? The influence of task specificity and prior knowledge on students' search strategies in hypertext. *Interacting with Computers, 15*, 409-428.

Rouet, J.-F., Favart, M., Britt, M. A., & Perfetti, C. A. (1997). Studying and using multiple documents in history: Effects of discipline expertise. *Cognition and Instruction, 15*(1), 85-106.

Ryan, M. P. (1984). Monitoring text comprehension: Individual differences in epistemological standards. *Journal of Educational Psychology, 76*(2), 248-258.

Schommer, M. (1990). Effects of beliefs about the nature of knowledge on comprehension. *Journal of Educational Psychology, 82*(3), 498-504.

Schommer, M. (1993). Epistemological development and academic performance among secondary students. *Journal of Educational Psychology, 85*(3), 406-411.

Schommer, M., Crouse, A., & Rhodes, N. (1992). Epistemological beliefs and mathematical text comprehension: Believing it is simple doesn't make it so. *Journal of Educational Psychology, 84*(4), 435-443.

Schommer-Aikins, M., & Hutter, R. (2002). Epistemological beliefs and thinking about everyday controversial issues. *The Journal of Psychology, 136*(1), 5-20.

Schraw, G. (1995). Measures of feeling-of-knowing accuracy: A new look at an old problem. *Applied Cognitive Psychology, 9*, 321-332.

Schraw, G., & Moshman, D. (1995). Metacognitive theories. *Educational Psychology Review, 7*(4), 351-371.

Stahl, E., & Bromme, R. (2007). CAEB. An instrument to measure connotative aspects of epistemological beliefs. *Learning and Instruction, 17*(6), 773-785.

Stahl, E., Pieschl, S., & Bromme, R. (2006).Task complexity, epistemological beliefs and metacognitive calibration: An exploratory study. *Journal of Educational Computing Research, 35*(4), 319-338.

Strømsø, H. I., Bråten, I., & Samuelstuen, M. S. (2008). Dimensions of topic-specific epistemological beliefs as predictors of multiple text understanding. *Learning and Instruction*, *18*(6), 513-527.

Tu, Y.-W., Shih, M., & Tsai, C.-C. (2008). Eighth graders' web searching strategies and outcomes : The role of task type, web experience and epistemological beliefs. *Computers & Education*, *51*(3), 1142-1153.

Weingardt, K. R., Leonesio, R. J., & Loftus, E. F. (1994). Viewing eyewitness research from a metacognitive perspective. In J. Metcalfe & A. P. Shimamura (Eds.), *Metacognition, knowing about knowing* (pp. 157-184). Cambridge: MIT Press.

Weinstock, M., & Cronin, M. A. (2003). The everyday production of knowledge: Individual differences in epistemological understanding and juror-reasoning skill. *Applied Cognitive Psychology*, 17, 161-181.

Windschitl, M., & Andre, T. (1998). Using computer simulations to enhance conceptual change: The roles of constructivist instruction and student epistemological beliefs. *Journal of Research in Science Teaching, 35*(2), 145-160.

Winne, P. H., & Hadwin, A. F. (1998). Studying as self-regulated learning. In D. J. Hacker, J. Dunlosky, & A. C. Graesser (Eds.), *Metacognition in educational theory and practice* (pp. 277-304). Mahwah, NJ: Lawrence Erlbaum Associates.

Winne, P. H., & Jamieson-Noel, D. (2003).Self-regulating studying by objectives for learning: Students' reports compared to a model. *Contemporary Educational Psychology*, 28, 259-276.

AFFILIATIONS AND ACKNOWLEDGEMENT

Rainer Bromme
Department of Psychology,
University of Muenster

StephaniePieschl
Department of Psychology,
University of Muenster

Elmar Stahl
University of Education, Freiburg

This study was funded by the German Research Foundation (DFG, BR 1126/3-1). We thank our student research assistants of this project (in alphabetical order): Stephanie Adrian, Sabine Joachim, Mike Pillukat, and TorstenPorsch. The data in this paper were collected and analyzed as partial fulfillment of the second author's dissertation requirements.

FLORIAN KLAPPROTH & PAULE SCHALTZ

THE VALIDITY OF PREDICTORS OF ACADEMIC AND VOCATIONAL-TRAINING ACHIEVEMENT: A REVIEW OF THE LITERATURE

INTRODUCTION

In many situations in educational settings, educational personnel are faced with problems that are associated with choosing the best option for a learner. For instance, teachers may ask about the age at which a child should enter primary school. In school systems with hierarchical tracking (e.g., in Germany, Switzerland, or Luxembourg), teachers may have to make a decision at the end of primary school about which track a student should attend in secondary school. Other teachers may have to decide which sets of instructions for single (or a group of) students will best fit the students' needs and capabilities. Universities may want to know whether applicants who have just passed their university entrance examination will eventually succeed at the university.

All these problems involve at least two different but consecutive evaluations of learners. First, the teacher (or someone else who is eligible) must recognise the actual abilities, wishes, preferences, or achievements of the learner. Thereafter, the teacher must gauge how these actual abilities etc. might develop in the future; that is, the teacher has to predict the future performance of a learner. A possible consequence of these two-fold evaluations is the selection of learners and their subsequent assignment to different learning environments that may foster the learner's abilities and may fit his or her educational or personal needs. However, educators or researchers may also be interested in examining the effectiveness with which achievements of learners at one stage of their educational careers can be predicted by achievements at an earlier stage.

Evaluating the accuracy with which learners' achievements can be predicted involves measuring predictive validity. Statistically, predictive validity is often estimated as the correlation between a predictor (i.e., a variable or a set of variables according to which the prediction is made) and a criterion (a variable or a set of variables whose amount is predicted by the predictor). The higher this correlation is, the closer is the relation between the predictor and the criterion. In cases of selection decisions, high predictive validity means that an individual (or a group of individuals) is likely to be selected correctly.

In the following, we present a review of studies that focussed on the quality of predictions made by educational personnel regarding future educational attainments of their students. The objective of this review is to provide a comprehensive overview of the state of the art in research that is concerned with the evaluation of the validity

S. Krolak-Schwerdt et al., (Eds.), Teachers' Professional Development, 153–166.

of predictions made at different points in time in the academic careers of students, in different educational contexts, and with different objectives. We categorised the studies that were included in our review according to the different stages of the educational system at which the predictions were made. In most countries, the first stage of institutionalised education usually begins with pre-school learning, which often occurs in kindergarten. In the next stage, children attend primary school, and after four to six (or more) years of learning in primary school, children move on to secondary school (stage three). In school systems with hierarchical tracking, the transition between primary school and secondary school necessitates decisions made by teachers about which track students should attend in secondary school. In the next stage of the educational system, some students go to university, whereas others take part in vocational training programmes or take up non-academic professions.

In order to obtain an exhaustive number of studies, we used the following databases for the literature search: *PsycINFO, PSYNDEX*, and *Educational Research Complete*. Keywords that were used for the literature search were: *validity, predictive validity, high school placement decision, transition from primary to secondary school, school achievement, school ability test, cognitive school ability test, school grades, college success, university entry test*, and their German equivalents. The search resulted in 211 studies. We included in our literature review those studies that fulfilled the following criteria: They were empirical studies, the studies reported results on the relation between at least one predictor and at least one criterion, data on the criterion were collected after the data were collected on the predictor, the context of the studies was educational in essence, and the studies were published before September 2012. After applying these criteria, a total of 52 studies remained for further analysis.

The review begins with studies that investigated the predictive validity of measures that were used for the prediction of students' achievements in primary school. Thereon, we reviewed studies that examined the validity of selection decisions or of accompanying measures, which were aimed at assigning primary school students to secondary school. Subsequently, we considered the prediction of achievements in universities and vocational-training programmes. Each section of the review is rounded off with a brief summary of the main results. At the end of this review, we discuss the main outcomes of this review. This overview should enable the reader (1) to evaluate how well academic achievements can be predicted on the basis of current measures in various educational domains, (2) to judge the effectiveness of selections in various educational domains in terms of assigning the "right" learner to the "right" condition, and (3) to assess the limits and constraints of predictions and selection decisions.

RESULTS

Prediction of Students' Achievements in Primary School

Results of the studies Studies that examined the relation between pre-school children's or first-graders' attributes and their achievements in primary school usually address one of two questions. They are either interested in predicting how well children will perform in primary school if information about some achievement indices from these children is already known; or they aim at identifying children who need some form of special education, for instance, students who are at risk of learning deficits or those who are talented.

The majority of studies assessing the predictive validity of pre-school achievements for later achievements in primary school used school-readiness tests as predictors and basic academic achievements in primary school as criteria. A child's readiness for school can be seen as a measurable set of pre-academic competences (including calculating, reading, drawing, and writing) and behavioural skills that have been demonstrated to predict later academic success (Augustyniak, Cook-Cottone, & Calabrese, 2004). Results of these studies provided moderate (r = .23 to .60) (Augustyniak et al., 2004; Baglici, Codding, & Tryon, 2010; Busch, 1980; Duncan & Rafter, 2005; Graue & Separd, 1989) to high (r = .73 to .84) (Jorgenson & Jorgenson, 1996; Swanson, Payne, & Jackson, 1981) validity coefficients regarding the relation between predictors and criteria. Although the studies differed regarding the period of prediction, from several months (Busch, 1980; Duncan & Rafter, 2005) to one year (Baglici, et al., 2010; Swanson, et al., 1981) to three years or more (Augustyniak et al., 2004; Jorgenson & Jorgenson, 1996), validity coefficients were similar across the different prediction periods.

La Paro and Pianta (2000) reported the results of a meta-analysis that examined the predictive validity of a variety of school-readiness screenings for later achievements in primary school. Their analyses allowed them to calculate effect sizes (which were averaged correlations, weighted by sample size). They found moderate effect sizes (mean r = .51) for the relation between pre-school and kindergarten achievements and between kindergarten and primary school achievements.

Moreover, children's motor skills and visual-motor coordination have been used as predictors for later achievements in primary school. The predictors that were used were test batteries including the assessment of motor skills (Funk, Sturner, & Green, 1986) or drawing tests (Haidkind, Kikas, Henno, & Peets; 2011). Both kinds of assessments yielded low to moderate (r = .09 to .70) validity coefficients.

Alongside cognitive abilities or motor skills, learning-related social skills have been used as criteria for students' success in primary school. Bart, Hajami, and Bar-Haim (2007) assessed the relation between basic motor abilities in kindergarten and scholastic, social, and emotional adaptation in the transition to formal schooling. The results indicated that motor functions showed significant predictive value for both scholastic and social/emotional adaptation to school (r = .23 to .58). Pagani, Fitzpatrick, and Parent (2012) examined the relation between children's kindergarten attention skills and patterns of classroom engagement throughout elementary school in disadvantaged urban neighbourhoods. Higher levels of kindergarten attention were associated with better classroom engagement (χ^2 (2) = 56.66).

Some studies investigated how accurately pre-school or early-school abilities are able to classify children who are in need of different instructional settings. For instance, Marx and Weber (2006) used a phonological processing task to identify children in kindergarten who were at risk for later reading and spelling deficits. For most criteria assessed in this study (spelling accuracy, reading speed, reading comprehension) in the four years of primary school, fewer than half of the children who developed later deficits were identified by the phonological processing task, and more than half of the children who were classified as at risk subsequently did not show a reading or a spelling deficit. On the contrary, Mazzocco and Thompson (2005) showed that learning deficits in mathematics in grades 2 and 3 of primary school were validly predicted by a composite of various cognitive tests. Logistic regression models correctly classified about 80% of the participants as having or not having learning deficits in mathematics. Similar results were obtained by Tröster, Flender, and Reineke (2011), who examined the predictive validity of a developmental screening test for kindergarten children.

Summary Most studies cited here revealed significant relations between pre-school achievements of learners and their subsequent achievements in primary school. This was true for different predictors (achievement tests, motor skills), different criteria (achievement tests, social skills), and across different periods of prediction (several months to four years). However, mixed results were obtained regarding the identification of special-needs students. A possible reason for the differences that were obtained may be due to the different degrees of similarity between the predictors and criteria employed within each study. Whereas some authors (e.g., Mazzocco & Thompson, 2005) used quite similar measures of the predictor and the criterion, others applied rather different measures (Marx & Weber, 2006). Similarity of measures (shared method variance) increases the correlation between these measures (cf. Lakin & Lohmann, 2011).

Prediction of Students' Achievements in Secondary School

Results of the studies In school systems with hierarchical tracking in secondary school (e.g., Austria, Germany, Luxembourg), school-placement decisions made by teachers at the end of primary school play a large role in determining the track a student will attend in secondary school. Because school-placement decisions imply a prediction of students' academic success in secondary school, the justification of school placement is achieved mainly by relating the performance of the students in secondary school to the initial assignment of the student to a certain track. A track is usually assumed to reflect the correct choice if the student who is initially assigned to that track remains in that track and exhibits sufficient performance. In the following, we will review studies that investigated the closeness of relations between predictors and criteria and studies that investigated how good the selections of students turned out to be for different educational settings on the basis of achievement or ability data that were obtained at an earlier point in time.

Spelberg and Rotteveel (1978) reported coefficients of correlation between scores on a language and math achievement test that was administered at the end of primary school and success in secondary school. These correlations were $r = .64$ on average. Within a large longitudinal analysis of students, Strand (2006) assessed the predictive validity of a national curriculum test and a cognitive-ability test at the end of primary school by relating the results of these tests to results of curriculum tests and examinations in secondary school (three to five years later). The cognitive ability test yielded slightly higher validity coefficients ($r = .63$ to .85) than the curriculum test ($r = .61$ to .81). However, a multiple regression analysis indicated that a combination of the two tests provided a better prediction of future outcomes. Thorsen (2012) reported a high predictive validity of school marks of ninth graders with respect to educational success in upper secondary school.

Recent studies on the predictive validity of school placement decisions stem from Roeder (1997), Schuchart and Weishaupt (2004), Scharenberg, Gröhlich, Guill, and Bos (2010), and Tiedemann and Billmann-Mahecha (2010). Roeder (1997) found that 81% of the students who attended the highest school track (Gymnasium) and were recommended for that track achieved success, whereas only 43% of those students who were recommended for lower tracks but nonetheless attended the highest track were effective. Similar results were obtained by Schuchart and Weishaupt (2004). They found that most of the students in the highest track who had been recommended for that track were successful (85%), but only a minority attending the same track but who had been recommended for a lower track succeeded (35%). The results from Scharenberg et al. (2010) were alike. 95% of the students in the highest track were successful if they had been recommended for that track, but only 67% succeeded without the corresponding school placement decision. Correspondingly, in a study conducted by Tiedemann and Billmann-Mahecha (2010), students were more likely to be successful if they attended a track they were previously assigned to (87%) than if they were not assigned to the track they actually attended (62%).

Sauer and Gamsjäger (1996) estimated the predictive validity of school placement decisions by using regression analysis. In this analysis, the school placement decision, several school marks, and scores from standardised aptitude tests served as predictor variables, and the criterion was school success in secondary school, measured by school marks. They found that school marks and test scores at the end of 4th grade in primary school were good predictors of school success in secondary school four years later, but the percentage of variance explained by school placement decisions was rather moderate (22%) when school marks and test scores were considered simultaneously. Similar results were obtained by Baeriswyl, Trautwein, Wandeler, and Lüdtke (2009).

Summary Students' achievements in secondary school could be well predicted by test scores obtained from the same students in primary school. Furthermore, there was a considerable relation between school placement decisions and school success.

Students who attended a recommended school track were more likely to succeed than students who chose the same track without a corresponding recommendation, although the number of students who successfully remained in a track when they were not recommended for that track was unexpectedly high. The overall small number of students who changed tracks also indicates that the school system is rather impermeable in general. Moreover, school placement decisions tend to predict school marks in secondary school to a significant degree. However, two findings may diminish the value of school placement decisions in predicting future school success. First, even for students who were in tracks for which they were not recommended, a remarkable percentage reached a sufficient level of achievement. A possible reason for the number of students who were successful in a track when they had not been assigned to it might be the fact that placement decisions are often affected by factors that are not related to students' academic abilities and skills (e.g., Ditton, Krüsken, & Schauenberg, 2005; Klapproth, Glock, Krolak-Schwerdt, Martin, & Böhmer, 2013). Second, when using school marks and test scores as additional predictor variables, the contribution of school placement decisions for the prediction of secondary school success was substantially smaller. This result seems to be plausible because school marks and test scores have been shown to be the most important determinants of school placement decisions (cf. Glock, Krolak-Schwerdt, Klapproth, & Böhmer, 2013).

Prediction of Students' Achievements in Universities and Vocational Training Programmes

Results of the studies Although school marks should reflect the academic competences of students, research shows that they are often contaminated by other factors such as the social status, age, or gender of the student, or other factors not related to achievement (cf. Baron-Boldt, Schuler, & Funke, 1988; Schuler, Funke, & Baron-Boldt, 1990). However, school marks are used prevalently as predictors of success in universities or vocational training programmes. The following section focusses on the predictive validity of school marks, different school achievement tests, and other selection tools used in the process of selecting secondary school students and undergraduates for further academic or vocational qualification.

It has been demonstrated by a variety of studies that school marks show moderate validity ($r \approx .40$) in predicting academic success in university programmes when the criterion for success is grade point average (GPA) (Passons, 1967; Schuler, Funke, & Baron-Boldt, 1990; Trapmann, Hell, Weigland, & Schuler, 2007). However, when study length is to be predicted, the predictive value of school marks is much weaker ($r = .28$; Höppel & Moser, 1993).

Additionally, scores on standardised tests have been used to predict academic success in universities. Most of them were scholastic-aptitude tests that yielded on average similar or even higher validity coefficients than school marks. De Sena and Weber (1965) found mean correlations of $r = .60$ between aptitude tests and GPA after two semesters. Gusset (1974) found correlations between *r* = .48 and *r* = .63 for the relation between aptitude test scores and grades in freshman mathematics. However, some studies have reported that school marks outperformed test scores in predicting academic success (Chrissom & Lanier, 1975; Geiser & Santelices, 2007; Stumpf & Stanley, 2002). The use of multiple predictors (e.g., school marks and test scores, or multiple tests) resulted in better predictions of academic success than single predictors, as has been found by Baron and Norman (1992) and Gumban and Iledan (1972).

Nevertheless, several factors have been found to affect the predictive validity of school marks and achievement tests. Trapmann et al. (2007) found that the predictive validity of school marks was moderated by the country of origin of the study, the study subject, and the length of time between when the school marks and university grades were obtained. Furthermore, Kobrin and Patterson (2011) found that the ethnic background of students affected the predictive validity of school marks and achievement test scores with respect to first-year college grades. Results found by Dalton (1976) indicated higher predictive validities of test scores for women (*r* = .49 to .64) than for men (*r* = .39 to .62).

In addition to investigating the predictive validity of some achievement indices in school, researchers and practitioners are interested in examining the effectiveness of tests or other procedures that students may have to take for admission to various universities and colleges. Predictors commonly used for this purpose are admission tests, assessment centres, and interview data. Compared to scholastic aptitude tests, admission tests are specialised to the needs of a certain college or programme (although some general aptitude tests may also serve as admission tests). Most studies were concerned with the predictive validity of medical school admission tests. The results indicated on average low (*r* ≈ .15) to medium (*r* ≈ .40) validity coefficients (Donon, Paolucci, & Violato, 2007; Emery & Bell, 2009; Poole, Shulruf, Rudland, & Wilkinson, 2012; Wilkinson, Zhang, & Parker, 2011). The combination of admission tests with school marks as predictors revealed higher validity coefficients than the use of a single test. Meagher, Lin, and Stellato (2006) found that a regression model that considered admission test scores and cumulative GPAs accounted for 21 to 37% of the variance between one to four years later, whereas a model that considered only admission test scores accounted for only 19 to 24%. Mitchel, Haynes, and Koenig (1994) found medium to high validity coefficients (*r* = .48 to .80) for a combination of GPAs and medical admission tests scores in predicting undergraduate grades, but when GPAs (*r* = .40 to .74) or test scores (*r* = .38 to .78) were considered alone for the prediction of undergraduate grades, the validity coefficients decreased. However, admission tests did not predict achievements in all subjects to the same degree (cf. Hell, Trapmann, Weigland, & Schuler, 2007; Stumpf & Fay, 1991). For example, Hell, Trapmann, and Schuler (2007) found different validity coefficients for human medicine (*r* = .51), veterinary medicine (*r* = .43), and dental medicine (*r* = .35). As Emery and Bell (2009) pointed out, one of the challenges of selecting students for medical school admission is that applicants generally all have equally high grades so that the main task is use their admission test scores to differentiate among them.

Bieri and Schuler (2011a,b) evaluated the predictive validity of an assessment centre that evaluated the cross-curricular competences of candidates in education. They found that students who successfully passed the tests at the assessment centre were more successful in their first year of study than those who performed low on the assessment centre's tests.

Although the predictive value of interviews is seen as controversial, interview data are widely used in the admission procedures of universities. In a meta-analysis, Hell et al. (2007) found a mean-corrected validity coefficient of $r = .16$, representing the correlation between admission interviews and university exam results. Similar results were found by Basco, Lancaster, Gilbert, Carey, and Blue (2008) and by Streyffeler, Altmaier, Kuperman, and Patrick (2005) for the prediction of clinical practice examination scores. There is some evidence that structured interviews ($r = .35$) outperform unstructured interviews ($r = .13$) in predicting medical examination tests (Eva, Reiter, Trinh, Wasi, Rosenfeld, & Norman, 2008; cf. also Hell et al., 2007).

Regarding the prediction of success in vocational training and qualification, Baron-Boldt, Schuler, and Funke (1988) found the predictive validity of school grades on average to be a little lower than for academic success ($r_{vocational} = .37$ versus $r_{academic} = .46$). A closer investigation focussing on different qualifications showed that the correlations varied between $r = .40$ for public administration and $r = .35$ for electronics. Schuler et al. (1990) found similar results and concluded that the accuracy of predictions regarding success in academic training was higher compared to the accuracy of predicting success in vocational training.

Summary The results presented above indicate that school marks and various school achievement tests have substantial predictive validity with regard to success at university or in vocational qualification. However, there are some differences regarding the amount of predictive value being worth mentioning. School marks and tests (including admission tests) seem to be better predictors of academic success than interview data. Academic success was better predicted than success in vocational training programmes. The use of multiple predictors yielded higher validity coefficients than the use of a single predictor. There are some factors that have been shown to moderate the amount of predictive validity. These factors consist of (among others) the student's major at university, the student's gender, the student's ethnicity, the lag between school marks and university grades, and the country of origin of the study.

DISCUSSION

This review presents four major results concerning the predictive validity of selection decisions and accompanying measures in educational contexts. To begin, one important result is that selection decisions and their accompanying measures have been shown to be valid in most of the cited studies as they have been shown to predict related outcomes to at least a moderate degree on average. This was true for all stages of the educational system and for all predictors and criteria used. Thus, we can safely state that the results of psycho-educational measurements are able to predict students' success at every stage within the educational system.

However, the second result obtained from this review is a limitation of the first one. Whereas there was overall validity in the prediction of individual success in school, universities, and training programmes, there were large differences in the amount of validity.

Differences in the amount of predictive validity were due to a variety of factors, of which the most important ones were the domain of prediction, the predictors used, and the prediction period. The highest indices of predictive validity were found in studies reporting on the selection of students for secondary school tracks. It was shown that most students who were selected for a certain track in secondary school remained in this track even some years later, whereas only a fraction of students had to abandon their track.

Another factor that contributes to differences in predictive validity is the predictor itself. Apparently, predictors that are measured in a highly standardised manner (for example, scores of standardised ability tests) turned out to correspond with predicted outcomes very closely, whereas rather unstandardised predictors (for example, job interviews) predicted outcomes to a smaller degree (e.g., Hell, Trapmann, Weigland, et al., 2007).

Furthermore, combining two or more predictor variables usually led to an increase in predictive validity (e.g., Baron & Norman, 1992). Moreover, validity coefficients were higher when the measurements of the predictor and criterion were more similar (Lakin & Lohmann, 2011).

Because predictions are made with regard to future outcomes, it is not surprising that the time interval between when the prediction or selection decision was made and when its validity was tested played a role in determining the amount of predictive validity. Actually, some studies reported here showed that coefficients of predictive validity were lower with a longer temporal distance between the measurements of the predictor and the criterion (e.g., Trapman et al., 2007). This decrease in predictive validity could be attributed to variability over time of the individual's abilities that are measured by the predictor variable (Althoff, 1984).

The third result is concerned with methodological problems of estimates of predictive validity. In educational systems with hierarchical tracks in secondary school (as in Germany or Luxembourg), tracking decisions are mainly based on students' achievements in primary school. Measuring the correctness of tracking decisions (hence, their predictive validity) is usually based on whether or not a student has remained in the track to which she or he had initially been assigned. According to this criterion, a high amount of predictive validity will be achieved if the quality of the tracking decision is high, that is, if teachers are able to validly predict students' future achievements in school. However, predictive validity will also be high if the school system's permeability is rather low. In the latter case, changing tracks is impeded because schools try to hold on to their students (regardless of their achievements). Thus, estimations of the predictive validity of tracking decisions may be biased if they are based on the number of students who remain in their original tracks after several years of schooling.

Limitations in interpreting predictive validity may also arise from the use of correlation or regression analyses to make estimates of predictive validity. Correlation or regression analyses often neglect non-selected individuals. For example, if school marks are considered as predictors of success at university, usually measures of the criterion of only those students who are selected for university are used, whereas students who are not selected are ignored. Because the assessment of results of criteria from non-selected individuals is hardly available, estimates of the predictive validity of selections are restricted to the selected sample and may even be biased compared to the whole sample.

The fourth result that we consider relevant for this review is related to evaluations of the quality of selection decisions. The predictive validity of selection decisions may be regarded as the degree to which selected individuals demonstrate that they are capable of adapting themselves and their achievement to the environment to which they were assigned. Correspondingly, the predictive validity of measures is the precision with which the results of these measures predict students' success at a certain stage of their academic career. However, even if predictive validity turns out to be high, a selection decision may nonetheless be disputable because the number of correctly or falsely classified individuals also depends on the base rate and the selection rate. The selection rate refers to the number of selected individuals divided by the number of all applicants, whereas the base rate is the number of "eligible" applicants divided by the number of all applicants. The selection rate often depends on the number of available places within a group or institution, for instance, the maximum number of students who are permitted to attend a course, whereas the base rate is unaffected by any selection strategy. It is clear that if the base rate is high and the selection rate is low, many applicants who are eligible would not pass the selection criterion, meaning that these individuals would be falsely classified. The predictive validity of the selection instrument is therefore a necessary yet not sufficient condition for the number of correctly and falsely classified individuals. Despite their acknowledged importance, base rates and selection rates are rarely considered for the interpretation of the amount of predictive validity. We think that in some studies that have reported coefficients of predictive validity, the neglect of base rates and selection rates could diminish the value of these studies.

CONCLUSION

Academic achievements in schools and universities as well as achievements in vocational training programmes can be predicted to a significant degree by using indices of achievements from a prior stage of education. In addition, the assignments of students to different courses or tracks, which are made on the basis of prior achievements, have been reported to be moderately or highly valid. However, particularly with respect to selecting students to secondary school tracks, high validity may reflect the school system's lack of permeability rather than teachers' precision in forecasting students' academic success. Several factors constrain the generalisability of the validity of predictors of academic and vocational-training achievement. These are the similarity of the predictor and criterion, the length of time between when the predictor and criterion are measured, restricting analyses to selected (and hence ignoring non-selected) individuals, and neglecting base rates and selection rates.

REFERENCES

Althoff, K. (1984). Zur prognostischen Validität von Intelligenz- und Leistungstests im Rahmen der Eignungsdiagnostik. *Predictive validity of intelligence and aptitude tests in aptitude testing. Psychologie und Praxis, 28*(4), 144–148.

*Augustyniak, K. M., Cook-Cottone, C. P., & Calabrese, N. (2004). The predictive validity of the Phelps Kindergarten Readiness Scale. *Psychology in the Schools, 41*, 509–516.

*Baeriswyl, F., Trautwein, U., Wandeler, C., & Lüdke, O. (2009). Wie gut prognostizieren subjektive Lehrerempfehlungen und schulische Testleistungen beim Übertritt die Mathematik und Deutschleistung in der Sekundarstufe I? *Zeitschrift für Erziehungswissenschaft, 12*, 352–372.

*Baglici, S. P., Codding, R., Tryon, G. (2010). Extending the research on the tests of Early Numeracy: Longitudinal analyses over two school years. *Assessment for Effective Intervention, 35*, 89–102.

*Baron, J., & Norman, M. F. (1992). SATs, achievement tests, and high-school class rank as predictors of college performance. *Educational and Psychological Measurement, 52*(4), 1047–1055.

*Baron-Boldt, J., Schuler, H., & Funke, U. (1988). Predictive validity of school grades: A meta-analysis. *Zeitschrift für Padagogische Psychologie, 2*(2), 79–90.

*Bart, O., Hajami, D., & Bar-Haim, Y. (2007). Predicting school adjustment from motor abilities in kindergarten. *Infant and Child Development, 16*(6), 597–615.

*Basco, W. T. Jr., Lancaster, C. J., Gilbert, G. E., Carey, M. E., & Blue, A. V. (2008). Medical School application interview score has limited predictive validity for performance on a fourth year clinical practice examination. *Advances in Health Sciences Education, 13*(2), 151–162.

*Bieri, C., & Schuler, P. (2011a). Check-point Assessment Centre für angehende Lehramtsstudierende. *Zeitschrift für Pädagogik, 57*(5), 695–710.

*Bieri, C., & Schuler, P. (2011b). Cross-Curricular competencies of student teachers: A Selection model based on assessment centre admission tests and study success after the first year of teacher training. *Assessment & Evaluation in Higher Education, 36*(4), 399–415.

*Busch, R. F. (1980). Predicting first-grade reading achievement. *Learning Disability Quarterly, 3*, 38–48.

*Chissom, B. S., & Lanier, D. (1975). Prediction of first quarter freshman GPA using SAT scores and high school grades. *Educational and Psychological Measurement, 35*(2), 461–463.

*Dalton, S. (1976). A decline in the predictive validity of the SAT and high school achievement. *Educational and Psychological Measurement, 36*(2), 445–448.

*De Sena, P. A., & Weber, L. A. (1965). The predictive validity of the School College Ability Test (SCAT) and the American College Test (ACT) at a liberal arts college for women. *Educational and Psychological Measurement, 25*(4), 1149–1151.

Ditton, H., Krüsken, J., & Schauenberg, M. (2005). Bildungsungleichheit – der Beitrag von Familie und Schule. *Zeitschrift für Erziehungswissenschaft, 8*, 285–304.

*Donnon, T., Paolucci, E. O., & Violato, C. (2007). The Predictive Validity of the MCAT for Medical School Performance and Medical Board Licensing Examinations: A Meta-Analysis of the Published Research. *Academic Medicine, 82*(1), 100–106.

*Duncan, J., & Rafter, E. M. (2005). Concurrent and predictive validity of the Phelps Kindergarten Readiness Scale-II. *Psychology in the Schools, 42*, 355–359.

*Emery, J. L., & Bell, J. F. (2009). The predictive validity of the BioMedical Admissions Test for preclinical examination performance. *Medical Education, 43*(6), 557–564.

*Eva, K. W., Reiter, H. I., Trinh, K., Wasi, P., Rosenfeld, J., & Norman, G. R. (2009). Predictive validity of the multiple mini-interview for selecting medical trainees. *Medical Education, 43*(8), 767–775.

*Funk, S. G., Sturner, R. A., & Green, J. A. (1986). Preschool prediction of early school performance: Relationship of McCarthy Scales of Children's Abilities prior to school entry to achievement in kindergarten, first, and second grades. *Journal of School Psychology, 24*(2), 181–194.

*Geiser, S., & Santelices, M. V. (2007). *Validity of High-School Grades in Predicting Student Success beyond the Freshman Year: High-School Record vs. Standardized Tests as Indicators of Four-Year College Outcomes.* Research & Occasional Paper Series: CSHE.6.07. Center for Studies in Higher Education. Retrieved from http://cshe.berkeley.edu/

Glock, S., Krolak-Schwerdt, S., Klapproth, F. & Böhmer, M. (2013). Prädiktoren der Schullaufbahnempfehlung für die Schulzweige des Sekundarbereichs I: Ein Überblick. *Pädagogische Rundschau*, *67*, 349–367.

*Graue, M. E., & Shepard, L. A. (1989). Predictive validity of the Gesell School Readiness Tests. *Early Childhood Research Quarterly, 4*(3), 303–315.

*Gumban, R. S., & Iledan, B. R. (1972). Some predictive validities of the College Entrance Test based on a relatively homogeneous group of schools. *Philippine Journal of Psychology, 5–6*, 16–42.

*Gussett, J. C. (1974). College Entrance Examination Board Scholastic Aptitude Test scores as a predictor for college freshman mathematics grades. *Educational and Psychological Measurement, 34*(4), 953–955.

*Haidkind, P., Kikas, E., Henno, H., & Peets, T. (2011). Controlled drawing observation for assessing a child's readiness for school and predicting academic achievement at the end of the first grade. *Scandinavian Journal of Educational Research, 55*(1), 61–78.

*Hell, B., Trapmann, S., & Schuler, H. (2007). Eine Metaanalyse der Validitaet von fachspezifischen Studierfaehigkeitstests im deutschsprachigen Raum. *Empirische Paedagogik, 21*(3), 251–270.

*Hell, B., Trapmann, S., Weigand, S., & Schuler, H. (2007). Die Validitaet von Auswahlgespraechen im Rahmen der Hochschulzulassung - eine Metaanalyse. *Psychologische Rundschau, 58*(2), 93–102.

*Höppel, D., & Moser, K. (1993). Die Prognostizierbarkeit von Studiennoten und Studiendauer durch Schulabschlussnoten. *Zeitschrift für Pädagogische Psychologie, 7*(1), 25–32.

*Jorgenson, C. B., & Jorgenson, D. E. (1996). Concurrent and predictive validity of an early childhood screening test. *International Journal of Neuroscience, 84*, 97–102.

Klapproth, F., Glock, S., Krolak-Schwerdt, S., Martin, R., & Böhmer, M. (2013). Prädiktoren der Sekundarschulempfehlung in Luxemburg: Ergebnisse einer Large-Scale-Untersuchung. *Zeitschrift für Erziehungswissenschaft*. doi:10.1007/s11618-013-0340-1

*Kobrin, J. L., & Patterson, B. F. (2011). Contextual Factors Associated with the Validity of SAT Scores and High School GPA for Predicting First-Year College Grades. *Educational Assessment, 16*(4), 207–226.

*La Paro, K. M., & Pianta, R. C. (2000). Predicting children's competence in the early school years: A meta-analytic review. *Review of Educational Research, 70*, 443–484.

Lakin, J. M., & Lohman, D. F. (2011). The Predictive Accuracy of Verbal, Quantitative, and Nonverbal Reasoning Tests: Consequences for Talent Identification and Program Diversity. *Journal for the Education of the Gifted, 34*, 595–623.

*Marx, P., & Weber, J. (2006). Kindergarten prediction of reading and spelling deficits: New results on the prognostic validity of the Bielefelder Screening (BISC). *Zeitschrift fur Padagogische Psychologie, 20*(4), 251–259.

*Mazzocco, M. M., & Thompson, R. E. (2005). Kindergarten Predictors of Math Learning Disability. *Learning Disabilities Research & Practice, 20*(3), 142–155.

*Meagher, D. G., Lin, A., & Stellato, C. P. (2006). A predictive validity study of the pharmacy college admission test. *American Journal of Pharmaceutical Education, 70*(3), 1–11.

*Mitchell, K., Haynes, R., & Koenig, J. (1994). Assessing the validity of the updated Medical College Admission Test. *Academic Medicine, 69*(5), 394–401.

*Pagani, L. S., Fitzpatrick, C., & Parent, S. (2012). Relating Kindergarten Attention to Subsequent Developmental Pathways of Classroom Engagement in Elementary School. *Journal of Abnormal Child Psychology, 40*, 715–725.

*Passons, W. R. (1967). Predictive validities of the ACT, SAT and high school grades for first semester GPA and freshman courses. *Educational and Psychological Measurement, 27*(4), 1143–1144.

*Poole, P., Shulruf, B., Rudland, J., & Wilkinson, T. (2012). Comparison of UMAT scores and GPA in prediction of performance in medical school: a national study. *Medical Education, 46*(2), 163–171.

*Roeder, P. M. (1997). Entwicklung vor, während und nach der Grundschulzeit: Literaturüberblick über den Einfluß der Grundschulzeit auf die Entwicklung in der Sekundarschule. In F.E. Weinert, & A. Helmke (Eds.), *Entwicklungen im Grundschulalter* (pp. 405–421). Weinheim: Beltz/PVU.

*Sauer, J., & Gamsjäger, E. (1996). *Ist Schulerfolg vorhersagbar? Die Determinanten der Grundschulleistung und ihr prognostischer Wert für den Sekundarschulerfolg.* Göttingen: Hogrefe.

*Scharenberg, K., Gröhlich, C., Guill, K., & Bos, W. (2010). Schulformwechsel und prognostische Validität der Schullaufbahnempfehlung in der Jahrgangsstufe 4. In W. Bos, & C. Gröhlich (Eds.), *KESS 8. Kompetenzen und Einstellungen von Schülerinnen und Schülern - Jahrgangsstufe 8* (pp. 115–123). Hamburg: Behörde für Schule und Berufsbildung.

*Schuchart, C., & Weishaupt, H. (2004). Die prognostische Qualität der Übergangsempfehlungen der niedersächsischen Orientierungsstufe. *Zeitschrift für Pädagogik, 50*, 882–902.

*Schuler, H., Funke, U., & Baron-Boldt, J. (1990). Predictive validity of school grades - A meta-analysis. Prognostische Validitaet von Schulabschlussnoten. *Applied Psychology: An International Review, 39*(1), 89–103.

*Spelberg, H. L., & Rotteveel, H. J. (1978). Predictive validity of the Groningen School Achievement Test. *Tijdschrift voor Onderwijsresearch, 3*, 3–9.

*Strand, S. (2006). Comparing the predictive validity of reasoning tests and national end of Key Stage 2 tests: which tests are the 'best'? *British Educational Research Journal, 32*, 209–225.

*Streyffeler, L., Altmaier, E. M., Kuperman, S., & Patrick, L. E., (2005). Development of a medical school admissions interview phase 2: Predictive validity of cognitive and non-cognitive attributes. *Medical Education Online, 10*, 1–5.

*Stumpf, H., & Fay, E. (1991). Zur prognostischen Validitaet des „Tests fuer medizinische Studiengaenge" (TMS). *den Studiengaengen Tier- und Zahnmedizin. Diagnostica, 37*(3), 213–225.

*Stumpf, H., & Stanley, J. C. (2002). Group data on high school grade point averages and scores on academic aptitude tests as predictors of institutional graduation rates. *Educational and Psychological Measurement, 62*(6), 1042–1052.

*Swanson, B. B., Payne, D. A., & Jackson, B. (1981). A predictive validity study of the metropolitan readiness test and meeting street school screening test against first grade metropolitan achievement test scores. *Educational and Psychological Measurement, 41*, 575–578.

*Trapmann, S., Hell, B., Weigand, S., & Schuler, H. (2007). Die Validität von Schulnoten zur Vorhersage des Studienerfolgs - eine Metaanalyse. *Zeitschrift für Pädagogische Psychologie, 21*(1), 11–27.

* Thorsen, C. (2012). Teachers' grade assignment and the predictive validity of criterion-referenced grades. *Educational Research & Evaluation, 18*, 153–172.

*Tiedemann, J., & Billmann-Mahecha, E. (2010). Wie erfolgreich sind Gymnasiasten ohne Gymnasialempfehlung? Die Kluft zwischen Schullaufbahnempfehlung und Schulformwahl der Eltern. *Zeitschrift für Erziehungswissenschaft, 13*, 649–660.

*Tröster, H., Flender, J., & Reineke, D. (2011). Prognostische Validität des Dortmunder Entwicklungsscreening für den Kindergarten (DESK 3-6). *Diagnostica, 57*(4), 201–211.

*Wilkinson, D., Zhang, J., & Parker, M. (2011). Predictive validity of the Undergraduate Medicine and Health Sciences Admission Test for medical students' academic performance. *The Medical Journal Of Australia, 194*(7), 341–344.

*References marked with an asterisk indicate studies that have been reviewed.

AFFILIATIONS AND ACKNOWLEDGEMENT

Florian Klapproth
University of Luxembourg

Paule Schaltz
University of Luxembourg

This study was supported by grant from the Luxembourgish Fonds National de la Recherche (Grant FNR/C11/LM/1201911).

INDEX