Rheinland-Pfälzische
Technische Universität
Kaiserslautern
Landau

TOWARDS A BETTER UNDERSTANDING OF TEACHER JUDGMENT ACCURACY: AN INTEGRATIVE APPROACH

Vom Promotionsausschuss des Fachbereichs 8: Psychologie
der Rheinland-Pfälzischen Technischen Universität Kaiserslautern-Landau
zur Verleihung des akademischen Grades Doktor der Philosophie (Dr. Phil.) genehmigte
Dissertation

vorgelegt von

Caroline Verena Bhowmik

Tag der wissenschaftlichen Aussprache: 28. November 2024
Vorsitzende des Promotionsausschusses: Prof. Dr. Tanja Lischetzke

Gutachter (Berichterstatter):
1. Dr. Friedrich-Wilhelm Schrader
Rheinland-Pfälzische Technische Universität Kaiserslautern-Landau
ehemals Universität Koblenz-Landau
2. Prof. Dr. Mitja Back
Universität Münster
3. Prof. Dr. Ingmar Hosenfeld
Rheinland-Pfälzische Technische Universität Kaiserslautern-Landau

To my mom.

TABLE OF CONTENTS

List of Tables……3

Abstract ……...……...4

Zusammenfassung……..6

Acknowledgements ……..……..8

1 Introduction……10

1.1 Teacher Judgment Accuracy and the Construct of Diagnostic Competence…….12

1.2 How Accurate are Teacher Judgments? - Evidence from Previous Research……...14

1.3 Which Factors can Foster (or Hinder) Accuracy of Teacher Judgments?.....17

1.4 Judgment Based on Minimal Information – Evidence and Methodological Approaches Rooted in Interpersonal Perception Research……...……..21

2 Aim of the Present Dissertation……..25

2.1 Manuscript 1: The Role of the Gender Bias for Teacher Judgments in the Physics Domain …….....……27

2.2 Manuscript 2: Teacher Judgments Based on Minimal Information Using Social Accuracy Analyses……..…….…..29

2.3 Manuscript 3: A Lens Model Approach to Teacher Judgment Accuracy ……...…….32

3 Discussion.....36

3.1 Summary……37

3.2 Implications of the Present Dissertation…….……...……..…..…38

3.3 Limitations and Directions for Future Research…….…...42

3.4 Conclusion……44

5 References…….……....46

Appendix……...……...68

Manuscript 1……..……..69

Manuscript 2…….....110

Manuscript 3…….123

Curriculum Vitae…….….156

Eidesstattliche Erklärung…….……...……..160

CRediT Author Statements……...162

LIST OF TABLES

Table 1. Short Summary of the Findings of the Present Dissertation…….…....38

ABSTRACT

Despite the relevance of accurate teacher judgments for successful learning processes and outcomes, previous research has revealed large differences between teachers in their ability to perceive students' characteristics accurately. Factors that could contribute to those differences in accuracy between teachers are to date not well understood.

Spanned over three individual research projects, this dissertation aims to provide a deeper insight into teacher judgment formation by approaching judgment accuracy from different theoretical and methodological angles. An integration of theoretical and methodological approaches that are regularly employed in social and personality psychology, i.e., interpersonal perception research, and that have not or rarely been employed in teacher judgment research, is proposed to generate a novel and more comprehensive picture of teacher judgment processes and outcomes.

In the first part of the present research (manuscript 1), the *gender bias* as possible factor to affect teacher judgments in the physics classroom was investigated. Findings showed that female students were underestimated in their academic self-concept in physics and perceived less accurately than their male classmates. In the second investigation (manuscript 2), judgments of students' intrinsic motivation, intelligence and academic self-concept were investigated based on brief videos (i.e., *thin-slices of behavior*) of unacquainted students (i.e., *zero-acquaintance approach*). Analyses based on the *Social Accuracy Model* (SAM; Biesanz, 2010, 2019) revealed the relevance of liking and attractiveness as moderators for (normative) teacher judgment accuracy. The final investigation (manuscript 3) demonstrates an application of *Brunswik's lens model* (BLM; Brunswik, 1956) to study the information (i.e., *cues*) that teachers use for their judgments, based on brief student videos, such as expressive behavior and physical appearance, and the validity of individual cues for the actual student characteristic (i.e., intrinsic motivation, intelligence and academic self-concept). Lens model parameter analyses revealed differences in cue validity across students' characteristics. Perceivers moreover relied on certain cues more often than on others, for instance students' sex and an attentive facial expression. In addition, a gender bias was detected as the boys in the sample overall received more favorable ratings than the girls. All investigations were carried out in the context of physics.

Taken together, leaning on and integrating theoretical and methodological perspectives from social and personality psychology has shown to bear a large potential for our understanding of teacher judgments and their accuracy. The outcomes of this dissertation may moreover serve as

orientational framework for the implementation of specific teacher training programs that support teachers' awareness for their perceptual processes.

ZUSAMMENFASSUNG

Urteilsakkuratheit, weithin bezeichnet als *Diagnostische Kompetenz*, wird als eine Kernkompetenz von Lehrkräften definiert. Studien zufolge unterscheiden sich Lehrkräfte in ihrer Fähigkeit, Schülermerkmale genau einzuschätzen. Untersuchungen zu möglichen Einflussfaktoren auf die Urteilsakkuratheit von Lehrkräften ergeben bisher jedoch ein uneinheitliches Bild. Ziel der vorliegenden Dissertation war es daher, auf der Basis dreier separater Untersuchungen unser Verständnis über jene Prozesse und Faktoren zu erweitern, die bei der Lehrerurteilsbildung eine Rolle spielen. Dabei wurde eine Integration theoretischer und methodologischer Ansätze aus der Sozial- und Persönlichkeitspsychologie angestrebt, die bisher hauptsächlich in der Forschung zur interpersonalen Wahrnehmung zum Einsatz gekommen sind und in jenem Bereich einen wichtigen Beitrag zum Verständnis zwischenmenschlicher Urteile geleistet haben.

Im ersten Teil der Arbeit (Artikel 1) wurde die Rolle von Geschlechtsunterschieden für Lehrerurteile im Sinne eines sog. *Gender Bias* untersucht. Es konnte aufgezeigt werden, dass weibliche Lernende hinsichtlich ihres Fähigkeitsselbstkonzepts unterschätzt und gleichzeitig weniger akkurat eingeschätzt wurden, als ihre männlichen Mitschüler. In der zweiten Untersuchung (Artikel 2) wurde die Akkuratheit von Lehrkrafturteilen hinsichtlich des akademischen Selbstkonzepts, intrinsischer Motivation und Intelligenz von Lernenden untersucht. Die Urteile wurden mithilfe kurzer Videoausschnitte von Lernenden, sog. *thin slices of behavior,* erhoben. Unter Anwendung des *Social Accuracy Modells* (SAM; Biesanz, 2010, 2019), wurde zudem die Rolle diverser Einflussfaktoren für die Urteilsakkuratheit überprüft. Ergebnisse zeigten, dass sowohl Sympathie, als auch körperliche Attraktivität für die (normative) Akkuratheit von Bedeutung sind. Die dritte Untersuchung demonstriert schließlich die Möglichkeiten der Anwendung von *Brunswik's Linsenmodell der Wahrnehmung* (BLM; Brunswik, 1956) im Rahmen der Lehrerurteilsforschung und für die Aufschlüsselung des Lehrerurteilsprozesses an sich. Somit wurde anhand kurzer Videoausschnitte untersucht, welche Hinweisreize (sog. *Cues*), d.h. Verhalten und Aussehen der Lernenden, für die Urteile herangezogen werden und welche Eigenschaften tatsächlich mit den untersuchten Kriterien der Lernenden (akademisches Selbstkonzept, intrinsiche Motivation und Intelligenz) einhergehen. Linsenmodellanalysen deuten auf Unterschiede hinsichtlich der Validität der Cues zwischen den untersuchten Kriterien hin. Im Urteilsprozess wurde darüber hinaus manchen Cues mehr Gewicht beigemessen, als anderen, wie beispielsweise dem Geschlecht der Lernenden und einem aufmerksamen Gesichtsausdruck.

Zudem konnten Hinweise auf einen Gender Bias ausgemacht werden. Alle Untersuchungen wurden im Rahmen des Physikunterrichts durchgeführt.

Zusammenfassend konnten die in der vorliegenden Dissertation vorgestellten Studien zeigen, dass eine Integration theoretischer und methodologischer Zugänge, sowohl aus der pädagogischen Psychologie und den Erziehungswissenschaften, als auch der Persönlichkeits- sowie Sozialpsychologie, gewinnbringend sein kann, da somit Aspekte der Lehrerurteilsbildung beachtet werden, die konventionelle Ansätze nicht ausreichend oder nur teilweise abzudecken vermögen. Zudem können die vorliegenden Ergebnisse als Grundlage für die Entwicklung von Trainings- und Weiterbildungsprogrammen für die Lehrerbildung dienen, die eine Sensibilisierung von Lehrkräften für ihre individuellen Urteilsbildungsprozesse zum Ziel haben.

ACKNOWLEGEMENTS

The present work would not have been possible without the guidance and the support of several individuals who in one way or the other played an important role for the success of this dissertation.

First and foremost, I would like to express my deepest appreciation to my advisors, Dr. Friedrich-Wilhelm Schrader and Prof. Mitja D. Back for their valuable and continuous support throughout my Ph.D. journey. Through his distinctive experience with teacher judgment accuracy research, Friedrich-Wilhelm Schrader positively contributed to my research on so many levels. He has always been approachable and taught me the importance to focus on the details in every step of the research process. I was moreover lucky to have Mitja Back on my team, who is exceptionally knowledgeable in person perception research and who inspired and motivated me from the moment we started collaborating on this research. Thank you, Mitja, for accompanying me on the way and for enriching my research with your honest and valuable advice. I am also highly thankful to Prof. Steffen Nestler for sharing his in-depth methodological knowledge with me and for his patience, considering that I did not have much of a statistical background when I started my Ph.D. I will surely benefit from this experience in my further scientific career.

I am also grateful to the German Research Foundation (DFG) for the financial support of my research through a Ph.D. scholarship at the Graduate School *Teaching and Learning Processes* (UPGRADE) at the Rheinland-Pfälzische Technische Universität Kaiserslautern-Landau (RPTU). Thank you for equipping me with the necessary means to conduct this research and for making the majority of my conference participations possible.

I owe a high level of gratitude to my student assistants Malika Reinmöller, Danielle Schmitt, Yvonne Dose, Sonja Fee Swidersky, Clarissa Gysen, Jan Rother, and Adrian Frech for their reliable commitment and identification with my research. This work would not have been possible without your support in terms of data collection, participant care, organization, and attentiveness in those intensive times when the studies were conducted.

My time in graduate school was particularly enjoyable due to so many wonderful colleagues - I would like to thank Prof. Anna-Katharina Praetorius for mentoring me and walking beside me until I was advanced enough to walk by myself. And my office mate and friend Katrin Hochdörffer. The team we formed was unique and enriching in so many ways! I am thankful also

to Dr. Loredana Mihalca, who always had an open ear, for helping me in finding my research direction, and who became a dear friend over time.

I am deeply grateful to Dr. Katja Dindar, whom I met during a summer school in Helsinki and with whom I developed a deep friendship over the years. Although miles apart, you have been on my side since then, sharing the challenges and joy of academia and life as such. Katja has become an indispensable part of my life. I am so grateful also for my other two friends for life, Anja and Tina, for being on my side for so many years and for providing me with support, strength and continuity throughout this journey.

I wish to express my heartfelt gratitude to my family, first of all my mother Claudia, whom I so dearly miss every single day, my father Bernd and my sister Vanessa for their encouragement and trust. I am highly grateful to my parents, who raised me with a passion for education and knowledge. Thank you for all your support, particularly in my university studies, knowing that this was related to countless sacrifices on your part. Thank you, Vanessa, for always being supportive and open to whichever idea is on my mind and for sharing every important moment of this and life's journey with me.

Finally, my sincere gratitude goes to my husband Avit, who is my soulmate, my travel buddy, my personal chef, my source of inspiration and positivity, and the best dad I could wish for to raise our wonderful daughters Ayana and Madita with. Avit has contributed to the success of this thesis both technically (by opening the world of R to me) and emotionally. Through his constant belief, I developed the necessary confidence that I can make it through every obstacle on the way and that I can be a successful scientist – no matter what. Thank you.

1 INTRODUCTION

"Seldom, very seldom, does complete truth belong to any human disclosure; seldom can it happen that something is not a little disguised or a little mistaken."

— Jane Austen, Emma

1 Introduction

Teacher judgments and their accuracy have been of interest for researchers within the discipline of educational psychology for several decades. They represent a research area that has been receiving increased attention over time given the growing evidence about the role and importance of teacher expectations and the accuracy of teacher judgments for educational processes and outcomes as well as students' self-concepts and motivation (Artelt, 2016; Bergold, 2023; de Boer et al., 2010; Helmke & Schrader, 1987; Kriegbaum et al., 2019; Möller et al., 2016; Pielmeier et al., 2018; Rubie-Davies et al., 2006; Szumski & Karwowski, 2019).

While we know that teachers, on average, can be fairly accurate in their judgments of students' academic performance and individual characteristics, such as the academic self-concept and intrinsic motivation, research has revealed substantial differences across individual teachers in their ability to form accurate judgments (Hoge & Coladarci, 1989; Lorenz & Artelt, 2009; Südkamp et al., 2012; Urhahne & Wijnia, 2021). However, possible influencing factors leading to those differences are to date not well understood, as evidence from previous work remains inconsistent (Südkamp et al., 2012; Urhahne & Wijnia, 2021). When analyzing influencing factors for teacher judgments, also teacher expectations as well as cognitive biases need to be considered (e.g., *gender bias*), which have shown to evoke perceptual distortions that can lead to inaccurate judgments across contexts and academic disciplines (Leslie et al., 2015; Murphy et al., 2003). These effects particularly concern STEM fields and subjects, such as for example physics, which have traditionally been dealing with gender-related issues of stereotyping and inequal dissemination of opportunities between girls and boys.

In light of the above, more research is needed to expand our knowledge about the nature of teacher judgment accuracy, which represents a crucial, yet complex and challenging task, both on the theoretical as well as the methodological level. Traditional approaches from educational science may therefore benefit from an integration of alternative methodological strategies, such as those that are widely acknowledged in social and personality psychology, where research about inter-individual perception and personality judgment have profoundly been studied in the last decades (e.g., Ambady & Rosenthal, 1992; Breil et al., 2021; Connelly & Ones, 2010; Karelaia & Hogarth, 2008; Kaufmann et al., 2013; Nestler & Back, 2013; Stopfer et al., 2014). Such approaches include the Social Accuracy Model (SAM; Biesanz, 2010, 2019) and Brunswik's Lens Model (BLM;

Brunswik, 1956), both of which have demonstrated different strengths: Whereas BLM highlights the importance of perceptual cues and the utilization of relevant cues by the perceiver for the judgment process (see Breil et al., 2021 for a recent overview), SAM enables the investigation of a wide range of possible moderators (see 1.3 and Funder, 1995, for a detailled elaboration of relevant moderator categories), while differentiating between normative and distinctive accuracy in a single analytic step. Research employing either of these models have oftentimes been combined with the *thin-slices of behavior* and *zero-acquaintance* approach (Albright et al., 1988; Ambady et al., 1995; Ambady & Rosenthal, 1992; Carney et al., 2007; Kenny & West, 2008; Murphy et al., 2014), two closely related research designs that bear great potential for application in teacher judgment accuracy research. They imply collecting judgments of previously unknown targets (i.e., zero-acquaintance) based on very brief information (i.e., thin-slices of behavior), such as pictures or videos, which allows for an exclusive investigation of (spontaneous) judgment formation processes independent of previous information or interaction between the perceiver and the target. Particularly for the case of teacher judgments, which are normally investigated following longer periods of previous teacher-student interaction and with plenty of information about the students available to the teachers, this research design allows to take a closer look at the teacher judgment process *per se*, including various potential moderating factors and the utilization of perceptual cues.

1.1 Teacher Judgment Accuracy and the Construct of Diagnostic Competence

The term *diagnostic competence* is predominantly used in the German speaking scientific context ("Diagnostische Kompetenz" Schrader, 1989; Spinath, 2005). In the English-speaking context sometimes referred to as *teachers' assessment competence* (e.g., Herppich et al., 2018), teachers' diagnostic competence is primarily dealt with in the context of two different meanings: (1) In its broadest sense, diagnostic competence is characterized as the ability to successfully perform regular diagnostic tasks in the educational context. This includes the selection, implementation, evaluation, and interpretation of suitable diagnostic procedures (e.g., different types of tests) and their utility for optimizing learning processes as well as for formal and informal evaluations of students. Teachers' diagnostic competence has received growing attention in connection with research on instructional quality and has progressively been regarded as integral part of the teaching profession, as reflected in the national educational standards in Germany

(Baumert & Kunter, 2006; KMK, 2022). As such, diagnostic competence has also become an integral part of teacher education curricula and textbooks in German higher education institutions. However, there is no general agreement or empirical evidence about the kind of ability and knowledge that is necessary to successfully perform these diagnostic tasks as well as the extent, to which these processes are conducted in the regular teaching practise (except for a few investigations concerning the general shortcomings of grading, e.g., Ingenkamp, 1972; Kronig & Ingenkamp, 2022). (2) In the narrower sense, diagnostic competence has been equated with *judgment accuracy*, even though accuracy as such only constitutes a facet of the broader construct. Judgment accuracy as the outcome of the judgment process can be regarded as a central aspect of diagnostic competence, as it refers to the quality of a teacher's diagnostic judgments. In educational research, as well as educational psychology, teacher judgment accuracy has extensively been studied since the 1980's including several meta-analyses and reviews and involving teacher judgments of different student characteristics, ranging from academic performance to intelligence or motivation ratings (Hoge & Coladarci, 1989; Kaufmann, 2000; Machts et al., 2016; Südkamp et al., 2012; Urhahne & Wijnia, 2021). When investigating teacher judgment accuracy, researchers have so far mainly been approaching the topic from two different theoretical angles: (1) Teacher judgments regarding students' characteristics (e.g., performance, academic self-concept or intrinsic motivation) as a prerequisite for successful adaptive teaching (Corno, 2008; Hardy et al., 2019; Hattie, 2009, 2012; Seidel & Shavelson, 2007) and (2) Teacher expectations and potential perceptual biases, such as for instance the gender bias, that represent influencing factors which can alter teacher judgments and their accuracy and can lead to *self-fulfilling prophecies* or *interpersonal expectancy effects* (Brophy, 1983; de Boer et al., 2010; Dusek & Joseph, 1983; Fiedler et al., 2002; Friedrich et al., 2015; Garcia et al., 2019; Jussim, 1989; Jussim & Harber, 2005; Rubie-Davies et al., 2006; Rubie-Davies, 2010; Timmermans & Rubie-Davies, 2018; Wang et al., 2018). In respect to adaptive teaching (1), the core assumption is that teacher judgments are crucial to the direction of classroom interaction, particularly for the adaptation of instruction, that is, adjusting one's teaching style to the individual needs and learning prerequisites of students. While *macro-adaptation* refers to teachers' decisions on a rather long-term basis, *micro-adaptations* refer to spontaneous reactions of teachers in accordance to situational requirements, that is, interactive decisions during classroom interaction, such as the insertion of an additional explanation, exercise or repetition opportunity, as well as the reaction to student mistakes. For the case of macro-

adaption, teachers plan their day to day lessons according to their perceptions of students' requirements, which, in turn, has an impact on various essential instructional decisions, such as grouping of students, and the selection of learning material (Schrader, 2013; Schrader & Helmke, 2014). Apart from the role of (accurate) teacher judgments for adaptive teaching, teacher expectations and perceptual biases (2) represent another core aspect of research concerning the accuracy of teacher judgments. Teacher expectancies and perceptual biases, such as the gender bias, can have undesirable, sometimes even harmful effects for students, both on the academic and on the individual level. Research on the formation of interpersonal judgments has shown that individuals tend to allocate more weight to the information that is learned first (also referred to as *primacy effect*; see works of Anderson & Barrios, 1961; Asch, 1946). Such first impressions about others can lead to expectancies about the other person that have shown to be stable over time (Harris & Garris, 2008). The tendency of perceivers to maintain the initial impression and interpret future information accordingly is referred to as the *confirmation bias* (Jussim & Harber, 2005; Nickerson, 1998). For the case of teacher judgments and expectations, first impressions can take effect immediately as well as over the course of several (school) years (Human et al., 2013), which is also described as the *teacher-student expectancy confirmation* (Darley & Fazio, 1980). Teacher expectations may moreover be expressed through teacher behavior in daily teacher-student interactions, particularly those involving teacher feedback (Wang et al., 2018). The transmission of expectations about individual students through teacher behavior has shown to influence students' self-concepts (e.g., Pesu et al., 2016; Upadyaya & Eccles, 2015) and motivation (e.g., Hornstra et al., 2018; Wang et al., 2018), which again can influence students' learning aspiration and academic outcomes.

1.2 How Accurate are Teacher Judgments? - Evidence From Previous Research

Teacher's judgment accuracy is normally assessed via teacher- or classroom-based correlations between the judgments and the corresponding standardized test scores or students' self-report data (for the case of non-cognitive criterion variables). This illustrative and comprehensive method is commonly referred to as *rank order component* (Schrader, 1989; Schrader & Helmke, 1987), as it reflects a teacher's ability to rank the students of their class or group according to their performance or indications on the assessed constructs (e.g., motivation or self-concept). The rank order component represents one out of three analytical components that

were proposed by Schrader and Helmke (1987) in their adaptation of Cronbach's (1955) rather complex component model to the educational context. Apart from the rank order component, the approach further distinguishes between a *level component* and a *component of differentiation*. In contrast to the rank order component, which comprises correlation based self-other agreement values between teachers and students, the level component addresses the difference between the average judgment and the average values of the students and is calculated by subtracting the mean of students' self-reports on a certain scale from the mean of teachers' judgments. For the case of the academic self-concept, positive values would indicate an overestimation and negative values an underestimation of the average level of students' self-concept (whilst the value "0" would indicate an ideal judgment). Finally, the component of differentiation depicts the agreement between teachers and students in terms of variation. This component therefore mirrors the tendency to over- or underestimate the variability of a certain characteristic in contrast to the actual variability. The component of differentiation is calculated by dividing the variation of teachers' judgments by the variation of students' self-reports or performance values. In this case, values larger than 1 would indicate an overestimation of the variability and values smaller than 1 an underestimation of the variability in the assessed group of students, whereas 1 would indicate an ideal judgment. Despite the added information that can be obtained from the individual accuracy components, the rank order component has been suggested to be the most relevant indicator of teachers' diagnostic competence (Schrader & Helmke, 1987).

Teachers' judgment accuracy has mainly been investigated in the school context and with a focus on judgments of students' academic achievement (for an overview, see Südkamp et al., 2012). This stems from the fact that performance related teacher judgments play a central role for many educational decisions, as well as the planning and direction of teaching and instruction in school. In this context, teachers are typically asked to provide their judgments about their students' expected results on a certain standardized test and subsequently, those judgments are compared with students' scores on these tests via correlational analyses. On average, teachers seem to be able to discern students' performance very well[1] (r = .63 across 75 studies; Südkamp et al., 2012 or r = .80 across 16 studies; Kaufmann, 2000). However, judgment accuracy outcomes have been shown to vary between subject areas and some teachers appear to be more accurate perceivers of students'

[1] To evaluate effect sizes, the benchmarks suggested by Funder & Ozer (2019) were employed throughout this dissertation as orientational framework.

academic achievement than others (e.g., $.11 \leq r \leq .88$; Lorenz & Artelt, 2009). In this context, possible causes of such accuracy differences across teachers have been discussed, a question that was taken up by Spinath (2005), who investigated whether or not the individual accuracy components outlined above may be allocated to an underlying, general (acquired) ability. However, to date, there is no empirical evidence supporting that teachers' diagnostic competence can be interpreted as a general ability that some teachers posess more than others given that no significant relationship could be established between the accuracy components (Karst, 2012; Schrader, 1989; Spinath 2005).

The existing research on teacher judgment accuracy has been extended both on a theoretical and a methodological level during the last two decades. Teacher judgment accuracy research is no longer restricted to teacher judgments of students' academic performance. Instead, a diverse set of student characteristics that are expected to play a relevant role for students' learning processes and outcomes have been included and investigated, among which are academic self-concept, intrinsic motivation, learning anxiety, creativity, giftedness, social skills, and cognitive abilities (see Urhahne & Wijnia 2021 for a recent review as well as for instance Machts et al., 2016; Spinath, 2005; Urhahne et al., 2011). Findings of these studies indicate that teachers are slightly less accurate in their judgments of students' non-cognitive characteristics and cognitive abilities compared to their judgments of students' academic performance. This discrepancy in accuracy is often explained by an increased observability of ability and achievement related variables, such as reading skills and vocabulary. Hence, cues relevant for the performance judgments that can guide teacher judgments are revealed in the classroom on a regular basis, e.g., when a student reads a text out loud or through the results of a written vocabulary test. On the contrary, affective-motivational variables may be less accessible by teachers during classroom interaction (see also Funder, 1995; Letzring et al., 2006).

As an easy to conduct statistical procedure, the rank order component reflects the most commonly applied method to determine teacher judgment accuracy. However, a weakness of this approach lies in its context dependency, particularly regarding the dispersion of student indications on self-report scales or test data. The extent of this dispersion can alter the difficulty of the judgment process. The rank component moreover typically refers to the perception of individual students. Hence, the capacity to accurately perceive the average distribution of certain characteristics or abilities in a group of individuals (e.g., school class), which can be regarded at

least as important, is left out of consideration. Hence, drawing on theoretical approaches and methods that have been shown to be relevant and regularly utilized in social and personality research to study interpersonal perception processes and the accuracy of personality judgments appears to represent a promising trajectory. Hence, two methodological approaches that have the potential to address the limitations of the rank order component outlined above are BLM and the SAM. BLM does not only allow to investigate the importance of various influencing factors (i.e., cues) for the judgment and the criterion, it also enables a further differentiation of the accuracy component (correlation between the judgments and the criterion) into three sub-constituents (ecological validity, consistency and sensitivity): (1) Ecological validity quantifies the role of the context, that is, how well the criterion can be predicted through a given set of cues. Hence, it represents the upper border of possible judgment accuracy when a perceiver is relying on the given set of cues. (2) Consistency assesses the degree to which the conjunction of the cues for the single targets is performed in a comparable way. (3) Sensitivity refers to the extent to which perceivers select cues that are valid and relevant for the criterion, that is, that are concurrent with their predictive capacity. This differentiation into more detailed parameters allows for an in-depth understanding of the rather simple rank order accuracy component. As a so-called profile-based approach, the SAM (as described in more detail in manuscript 2) assumes that perceivers usually judge other individuals on a set or profile of different characteristics as opposed to only one. It also differentiates between two parameters: distinctive accuracy as individual-related accuracy, and normative accuracy as group-related accuracy parameter. These parameters are derived from a multilevel regression model, which moreover allows to include different moderators that can lead to high or low judgment accuracy outcomes among teachers.

1.3 Which Factors can Foster (or Hinder) Accuracy of Teacher Judgments?

Differences across individual teachers in their judgmental capbabilities have been demonstrated throughout the existing research (see recent review by Urhahne & Wijnia 2021), which raises the question regarding possible influencing factors (i.e., moderators) for teacher judgment accuracy. In the framework of their meta-analysis investigating teacher judgment accuracy in the context of achievement, Südkamp and colleagues (2012) also systematized different types of moderators for teacher judgment accuracy. These categories involve different characteristics of the judgment situation that either refer to the teacher (e.g., expectations or

teaching experience), the student (e.g., prior knowledge or motivation), the test, based on which students' achievement was measured (e.g., standardized vs. non-standardized measures or domain specificity) or the judgment task itself (e.g., informed vs. uninformed judgments). However, this analysis is confined to research on teacher judgments of students' achievement and only included studies applying the traditional correlational approach to teacher judgment accuracy (i.e., rank order component).

To identify informative approaches for investigating relevant moderators of teacher judgment accuracy, theories and concepts rooted in social and personality psychology, i.e., interpersonal perception and human judgment and decision making, have been found to offer valuable insights. As central conceptualizations of accurate judgment processes, BLM (Brunswik, 1956) and the Realistic Accuracy Model (RAM; Funder, 1995) therefore serve as theoretical framework for the present dissertation. This framework is embedded in a larger theoretical context, the *Social Judgment Theory*, which evolved in the 1950s and 1960s from Brunswik's probabilistic functionalism psychology and is represented in BLM (for further elaboration of the theory, see manuscript 2 and Cooksey, 1996; Doherty & Kurz, 1996; Hammond et al., 1986). Based on the lens model approach, RAM distinguishes between four stages in the judgment process that need to be met in order to achieve judgment accuracy. The first two stages primarily lie within the target itself and refer to (1) the presence of relevant (behavioral and physical) cues and (2) the availability of these cues to the perceiver. The following and last two stages are more of concern to the judge, which is (3) the detection of these cues and finally (4) the utilization of relevant cues for the judgment. Against this background, Funder (1995) argues that four categories of moderators determine how successful a judgment process can be navigated by a perceiver, that is *the good judge, the good target, the good trait and the good information*. Based on these categories, several moderators have been investigated in interpersonal perception research that may explain inter-individual differences in accuracy outcomes, such as a judge's intelligence or degree of interpersonal experience (Letzring, 2008), or target's psychological adjustment, social status, and socialization (Human & Biesanz, 2011; Human & Biesanz, 2013). Beyond judge and target, also the nature of the criterion itself has shown to lead to more or less accurate judgments. This phenomenon is often explained by the degree of observability of certain traits versus others, that is, how visible a trait is to a perceiver (Funder & Dobroth, 1987). The last category of moderators refers to the amount and quality of information, which is provided to the judge as a basis for the

judgment (Blackman & Funder, 1998; Letzring et al., 2006). It may for example play a role, whether the stimuli for a judgment are presented in a verbal, nonverbal or combined mode. According to RAM, information is of high quality, when it is available to the judge and relevant to the characteristic that is being judged. The categories of moderators outlined above can take effect individually, but also through interactions between separate moderating categories, such as judge-target interactions in terms of personality or interpersonal appeal (i.e., *relationship* or *dyadic moderators*).

In view of the foregoing, also moderators that have been investigated in teacher judgment accuracy research can be allocated to different categories, such as characteristics of the teacher, student, test, and judgment (Südkamp et al., 2012; Urhahne & Wijnia, 2021). Teacher characteristics that have been considered to be relevant for the accuracy of teachers' judgments include teaching experience (e.g., Ready & Wright, 2011), intelligence (e.g., Kaiser et al., 2012), judgment confidence (Praetorius et al., 2013), and familarity with the class or assessment task (in the case of performance judgments; e.g., Begeny & Buchanan, 2010; Oerke et al., 2015). Whereas teaching experience does not seem to foster teacher judgment accuracy, intelligence and judgment confidence, on the other hand, may indeed result in more accurate teacher judgments (Kaiser et al., 2012; Praetorius et al., 2013). Above and beyond characteristics of the teacher, potentially relevant student characteristics involve students' ability and language proficiency, both of which have shown to promote teacher judgment accuracy (Begeny et al., 2008, 2011; Eckert et al., 2006; Wijnia et al., 2016). Notably, such categories of moderators can take affect separately, or interact with each other (Funder, 1995; Südkamp et al., 2012). However, in contrast to the individual moderator categories, less evidence exists about the role of such relationship-specific factors in teacher-student dyads (dyadic moderators), which may foster or hinder judgment accuracy, such as for instance liking, attractiveness and personality similarity. The consideration of dyadic factors appears to be particularly relevant given that they have been shown to be robust over time (Fultz et al., 2023). In the educational setting, some evidence exists, that student attractiveness, as well as other non-cognitive factors, such as attentiveness and neatness influence teachers' judgments, and consequently teachers' expectancies towards the students (Doherty & Conolly, 1985; Dusek & Joseph, 1983; Parks & Kennedy, 2007; Ritts et al., 1992). Consequently, attractive students have shown to be evaluated more positively and a teacher's liking of a student has shown to be associated with higher performance in school (Murphy et al., 1981; Ritts et al., 1992). In other contexts,

attractive individuals have shown to receive more help (Dovidio et al., 2006) and to be evaluated as more intelligent than unattractive individuals (Feingold, 1992). Another moderator related to the teacher-student relationship is personality similarity between perceivers and targets. Research assessing similarity effects typically differentiates between the role of judges' assumed similarity to the target, often referred to as the degree of perceiving one's own characteristics in others (Cronbach, 1955) and actual similarity, that is, the degree of actual personality similarity between judge and target. Some studies have proven the positive affect of personality similarity on the judgment outcome (e.g., Chaplin & Panter, 1993; Funder et al., 1995; Letzring, 2010). However, in some cases, personality similarity may also lead to an increased liking of a target, which in turn can result in an overly positive and thus less accurate rating (Montoya & Horton, 2013). Findings of a study investigating teacher-student similarity effects revealed that students who were more similar to their teachers in respect to personality were judged more positively in their ability than students that were less similar (Rausch et al., 2015). A relationship moderator that has been investigated with more frequency is the gender bias and the role of potential gender stereotypes in teacher judgments (Bennett et al., 1993; Bonefeld et al., 2020; Garcia et al., 2019; Holder & Kessels, 2017; Muntoni & Retelsdorf, 2018; Ready & Wright, 2011; Timmermans et al., 2015). A gender bias describes the tendency to behave differently toward one gender compared to the other, often in favor of men or boys (Rothchild, 2007). In the school context, the *gender grading bias* (also: gender grading gap) has been receiving substantial attention, indicating differences in grades based on students' gender, even when the levels of academic performance are equivalent (e.g., Doornkamp et al., 2022; Protivínský & Münich, 2018). Such gender stereotypes have shown to be present across disciplines, but are particularly common in STEM fields, such as for example the physics classroom (e.g., Hand et al., 2017; Hofer, 2015; Leslie et al., 2015). Another example for a dyadic moderator exemplifying a perceptual bias are teachers' perceptions or expectations towards a student's ethnic background, indicating advantages of students belonging to an ethnic majority compared to those belonging to an ethnic minority (e.g., Glock, 2016). Moderating variables concerning the characteristics of the specific (achievement) test or the characteristics of the judgment task have also shown to be of relevance for teacher judgment accuracy, but are beyond the scope of the present dissertation.

Given that most of the available research was focused on teacher judgments of students' academic achievement and after longer periods of interaction between teacher and students, it is

unclear to what extend these findings are also true for the judgment of emotional-motivational student characteristics (e.g. students' academic self-concept and intrinsic motivation) and in situations of minimal information and no prior acquaintance between teachers and students.

1.4 Judgments based on minimal information – evidence and methodological approaches rooted in interpersonal perception research

Teacher judgment accuracy research has predominantly been conducted after longer periods of interaction between students and teachers in the regular classroom context. This entails that teachers can draw on previous knowledge or information about the students for their judgments, which can make it difficult to investgate the judgment process per se. To eliminate the knowledge that teachers have acquired after longer periods of interaction with the students and thereby being able to investigate processes involved in teacher judgment formation, a research design based on the thin-slices of behavior (Ambady & Rosenthal 1992) and/or zero-acquaintance (Albright et al., 1988) approach can be considered. These approaches have been proven to be well suited to study first impressions as a part of interpersonal perception and judgment accuracy research, which represents a large research field within social and in particular personality psychology (e.g., Ambady et al., 1995, 2000; Ambady & Rosenthal, 1992; Borkenau et al., 2004; Carney et al., 2007; Hirschmüller et al., 2013; Murphy et al., 2014; Murphy & Hall, 2021; Nestler & Back, 2013; Ambady et al., 2000). Both the thin-slices of behavior and zero-acquaintance methodologies are closely related and the terms are therefore typically used interchangeably (Jiang et al., 2023; Kenny & West, 2008). In a broader sense, they describe research designs in which lay perceivers form their judgments about other individuals they do not know based on brief videos or photographs, but also audio material, direct interactions or writing style (Back & Nestler, 2016). Hence, a common premise of both approaches entails that the perceivers do not interact with the targets prior to the judgment situation. However, Kenny and West (2008) suggest that in zero-acquaintance studies, perceivers and targets sometimes interact face-to-face, which can involve some kind of dyadic interplay between perceiver and target. This is not the case for thin-slice studies, in which the information available to the perceivers is rather reduced and judgments are solely performed based on brief video clips that serve as stimulus material. Such brief video clips usually range from 1 second to 5 minutes and are extracted from longer sequences (so-called "streams") of (oftentimes nonverbal) expressive behavior (Ambady et al., 2000; Ambady &

Rosenthal, 1992; Murphy et al., 2019; Murphy & Hall, 2021). Essentially, expressive behavior can involve all possible channels of communication, such as speech, voice, face and body, however, researchers conducting first impression and thin-slice research often rely on nonverbal target behavior (Ambady et al., 2000). This stems from the assumption that nonverbal expressions hold some unique features compared to vocal expressions, which can be beneficial to first impression formation processes. Thus, a relevant feature of nonverbal behavior is that it is mainly spontaneous and less controllable and accessible by the target, leading to a comparatively valid source of information (Ambady et al., 2000; DePaulo, 1992). In sum, the thin-slice approach is based on two assumptions: (1) personality is generally expressed and can be observed from behavior (Allport, 1937; Letzring et al., 2021). Hence, brief insights into a target's behavior may unveil relevant aspects of a target's personality characteristics as well as inner states, interaction motives, and social relations (Ambady et al., 2000; Murphy et al., 2014). And (2) behavior that is displayed in brief video snippets serves as a representative for a target's general behavioral pattern (Murphy & Hall, 2021).The second assumption stems from the *behavioral consistency premise*, a fundamental theoretical principle within personality psychology suggesting that an individual's behavior can be regarded as stable over time and across contexts (Funder & Colvin, 1991; Geukes et al., 2017; Murphy & Hall, 2021; Shoda, 1999). Hence, thin-slices can be regarded as valid and suitable means to study perceiver judgments regarding various target characteristics. The application of brief videos to study interpersonal perception has shown to be associated with several other advantages, one of which is the optimal utilization of available ressources, for instance by reducing the amount of time needed for extensive behavior coding and measurement. Moreover, this method reduces the necessary amount of participants given that one participant can rate several targets in a row when brief videos are employed (Murphy et al., 2014). Behavior measurements also represent a core element of BLM, which can be supported by employing thin-slices of behavior. The lens model thereby offers further insights into how certain measured target characteristics are expressed through behavioral cues while at the same time revealing perceivers' possible implicit theories and expectations about the characteristic or trait of interest (Murphy & Hall, 2021). Taken together, thin-slices of behavior can be regarded as suitable tool for studying interpersonal perception, not only due to practical considerations, but also given their demonstrated theoretical validity and relevance.

Evidence from previous thin-slice and zero-acquaintance research has shown that individuals make inferences of other individuals' characteristics across social situations quickly and with a substantial amount of accuracy (Ambady et al., 2000; Ambady & Rosenthal, 1992; Carney et al., 2007). Characteristics that are most often investigated include the Big Five of personality (John et al., 2008), i.e., openness, conscientiousness, extraversion, agreeableness, and neuroticisim, and intelligence, but also other characteristics have been of interest, such as self-esteem and attitudes (Beer & Watson, 2010; Kilianski, 2008). Accuracy levels are typically lying between .20 and .40 (aggregated perceiver analyses) and between .10 and .30 (single perceiver analyses) across various characteristics and social contexts (Back & Nestler, 2016). Whereas to obtain aggregated perceiver accuracy, the relationship between perceiver judgments and target values are calculated and averaged across perceivers for each trait (*variable-centered approach*), single perceiver accuracy is calculated separately for each perceiver across several traits a mean is calculated (*person-centered approach*) (Back & Nestler, 2016).

The application of the thin-slice or zero-acquaintance approach in educational research has to date been very limited. The method was first applied to the teaching context towards the end of the last century and confined to teacher expectations and biases (Babad et al., 1989a, 1989b), as well as teaching effectiveness (Ambady & Rosenthal, 1993). In one of the studies by Babad and colleagues (1989b), brief videos showed teachers when they were *talking about* vs. *talking to* students they were having strong and low expectancies for. In sum, teachers attempted to compensate their negative expectations through the more controllable behaviors and ways of communication, such as direct conversations with the students. However, in the less controllable, mainly nonverbal communication channels, such as facial expression, negative affect was transmitted nevertheless. Comparable results were obtained in another study by Babad and colleagues (1989a), in which brief videos of teachers with an identified bias as well as videos showing unbiased teachers (here referred to as the evaluation of products made by students from different religious backgrounds and/or ethnicity) were analysed. Results showed that biased teachers communicated in an even warmer fashion with their students than teachers with no identified bias. Notably, also here, teachers' biases were revealed nevertheless through their non-verbal expressive behavior in the brief video clips. The phenomenon, in which negative affect "leaks through" channels that are less controllable by individuals, is often referred to as the *leakage effect* (Babad et al., 1989a). The findings of both studies underpin the suggestion that the thin-slice

method may be very suitable to unveil important information, some of which may not even be identified in longer phases of interaction (Ambady et al., 2000). More recent work adresses the predictive validity of thin-slices for the assessment of instructional quality (Begrich et al., 2021) and for teachers' well-being (Pretsch et al., 2013). Given the methodological limitations of the traditional approach to study teacher judgment accuracy and judgment processes outlined in 1.2, the thin-slice and zero-acquaintance approach has recently also gained attention within teacher judgment accuracy research, however, only two studies have been conducted beyond the research presented within the present dissertation. In those studies, brief videos of previously unknown students were employed and teachers were asked to to indicate their judgments regarding students' social status and behavior (Lansu & Berg, 2020) as well as self-concept (Praetorius et al., 2015). The outcomes showed that teachers, just as any other human perceivers, can depict various student characteristics from thin slices of behavior and that accuracy outcomes may not benefit from longer periods of teacher-student interaction (Praetorius et al., 2015).

Drawing on the theoretical and practical relevance of the the thin-slice and zero-acquaintance approach, and the fact that it has repeatedly shown to be fruitful in exploring interpersonal judgments in interpersonal perception research, it appears to offer great opportunities also for research on teacher judgment accuracy. In particular, the main advantage over previous approaches can be allocated to the possibility to investigate teacher perceptual processes while keeping other influencing factors, such as information about students' social background or previous behavior, constant. Generally speaking, applying the thin-slice approach in teacher judgment accuracy research goes along with a shift from knowledge- and experience-based teacher judgments to teachers' first impressions, which then form the basis for subsequent expectations and judgments (Ambady et al., 2000).

2 AIM OF THE PRESENT DISSERTATION

Aim of the Present Dissertation

The present dissertation sought to identify factors that contribute to rather accurate or inaccurate teacher judgments regarding different student characteristics, including academic self-concept, intrinsic motivation, intelligence, estimate of effort, and perceived difficulty. One central aim of the present dissertation was therefore to cover a broader range of moderating factors to extend our knowledge about teacher judgment formation and accuracy. These factors included student- and perceiver-related moderators, such as for instance students' sex and teaching experience, as well as moderators of the teacher-student relationship, such as personality similarity, perceived attractiveness and liking. Of particular interest were moreover the investigastion of the role of perceptual cues (i.e., students' expressive behavior and physical appearance) and teachers' utilization of valid cues in their judgments.

To obtain this aim, the present dissertation encompasses three separate research projects, all of which were carried out in the physics context. The first investigation (manuscript 1) aimed at identifying the role of students' sex for teacher judgments as well as potential cognitive biases, such as the gender bias, in two individual sub-studies. In the second investigation (manuscript 2), a profile-based approach to teacher judgment accuracy was carried out by applying the Social Accuracy Model (SAM) with the goal to identify the role of different moderators for judgment accuracy. Drawing on Brunswik's Lens Model (BLM), the final part of the present research (manuscript 3) explored the differing roles of certain behavioral and physical cues that are either valid for the criterion of interest (i.e., students' characteristics) and/or the perceiver judgments. Perceiver judgments carried out in the research comprising manuscript 2 and 3 were based on brief student videos (thin-slices of behavior) who were unaquainted to the perceivers (zero-acquaintance).

The overarching aim of the research carried out in the present dissertation can be allocated to the moderating categories "teacher and student characteristics" as introduced in the meta-analysis on teacher judgment accuracy by Südkamp et al. (2012), while adding a category that focuses on moderators concerning the teacher-student relationship based on the Realistic Accuracy Model (RAM) by Funder (1995). By considering theoretical and methodological approaches that have been established and widely applied to study the accuracy of personality judgments in

psychological research, the present dissertation in large parts represents an integration of social/personality psychology and educational sciences.

2.1 Manuscript 1: The Role of the Gender Bias for Teacher Judgments in the Physics Domain

As it is the case for any situation involving interactions between individuals, diagnostic judgments in the school context can be distorted by expectation biases that can lead to inaccurate judgments and consequently (negatively) determine a teacher's behavior towards individual students (Boer et al., 2010; Cate et al., 2014; Glock et al., 2013; Parks & Kennedy, 2007; Szumski & Karwowski, 2019a; Wang et al., 2018). In respect to Science, technology, engineering, and mathematics (STEM) fields, which also includes physics education, the gender bias serves as a common explanation for a systematic cognitive distortion often resulting in inaccurate judgments as well as over- and underestimations, which can give rise to disadvantages mainly for female students (Hand et al., 2017; Hofer, 2015; Leaper & Starr, 2019; Wang & Degol, 2017). The majority of previous research on teacher judgments of students' performance has been conducted in the mathematics classroom (Jussim & Eccles, 1992; Robinson-Cimpian et al., 2014), however, gender biases appear to be equally prevalent in the physics context, resulting in lower grades for girls (Hofer, 2015) and favorable ratings of boys' performance orientation (Heller et al., 2001). These gender-biased expectations could be related to the fact that physics has traditionally been regarded as a rather *male* subject, where boys are expected to perform better than their same-aged female classmates (Graves et al., 2017; Kessels et al., 2008; Solga & Pfahl, 2009) and in which girls therefore continue to be underrepresented at all levels of the educational sector (Kost-Smith et al., 2010; Stewart, 1998; Zohar & Bronshtein, 2005). In contrast, research comparing actual academic outcomes between boys and girls in the natural sciences has shown that girls often outperform boys (Herwartz-Emden et al., 2012), but at the same time tend to indicate lower self-concepts (Jansen et al., 2019; Kessels et al., 2008). Other studies revealed lower performance rates for female students in physics (e.g., Kost-Smith et al., 2010; Reiss et al., 2016), which among other factors could be related to the fact that girls tend to receive lower grades in physics than their male classmates (Hofer, 2015). Hence, most existing research dealing with gender-related teacher expectations in STEM fields has been focused on achievement ratings. Given the relevance of teacher judgments regarding students' emotional and motivational characteristics for students' academic prospects (e.g., Artelt, 2016; Bergold, 2023; de Boer et al., 2010), investigating gender

in respect to these characteristics appears to be equally relevant. Therefore, the two studies comprising manuscript 1 were focused on teacher judgments of students' academic self-concept alongside two metacognitive constructs, i.e., feeling of difficulty and estimate of effort. Previous research mainly considered the subject-independent, broad academic self-concept, despite evidence from previous research underlining the domain-specifity of the construct (Möller et al., 2016). The present studies extend this line of research by investigating teacher judgments of students' academic self-concept in the context of physics. In more general terms, a person's self-concept is generally referred to as knowledge about oneself, including one's performance and ability and is developed and influenced through interactions with the environment (Shavelson et al., 1976). In the academic context, students' views on their own competencies, as well as teachers' perceptions thereof, have been shown to be a predictor for students' academic achievement (Huang, 2011; Marsh, 1992; Marsh & Craven, 2006; Möller et al., 2016; Praetorius et al., 2016; Schrader & Helmke, 2015; Wigfield & Karpathian, 1991). As a central motivational characteristic, students' academic self-concept can determine the amount of effort that is invested by a student when solving given tasks as well as the level of perceived task difficulty and vice versa (Dapp & Roebers, 2021; Efklides, 2006; Efklides & Tsiora, 2002). Estimates of effort together with feelings of difficulty are part of *metacognitive experiences*, which represent further facets relevant to the learning process that are potentially prone to teacher perceptual biases, particularly in the physics context. Metacognitive experiences have mainly been investigated in research on students' self-regulatory competence (Efklides, 2011, 2017; Efklides & Tsiora, 2002; Flavell, 1979; Veenman et al., 2006; Zimmerman, 1995). As a part of the cognitive monitoring process in learning settings, students' ability to monitor their estimate of effort and feeling of difficulty are regarded as important prerequisite for successfully completing a given task (Efklides, 2017).

Research comprising the first manuscript within this dissertation was carried out by implementing two separate studies that were designed to investigate the role of students' ($N = 207$) sex for the perceptions of $N = 27$ pre-service teachers (study 1) and $N = 10$ experienced teachers (study 2) in the physics classroom. Students participating as targets in these studies were in 9th grade with an average age of $M = 15.6$ Jahre ($SD = 0.63$) and 40.3 % were female. In study 1, pre-service teachers indicated their judgments of previously unknown students (on average eight students per teacher) after they conducted a 90-minute long teaching episode with those students. In study 2, experienced teachers indicated their judgments for the students of their own classrooms

(on average 21 students). The main aim was to identify a possible gender bias in the judgments of students' academic self-concept in physics and their metacognitive experiences (i.e., feeling of difficulty and estimate of effort). In respect to the rank order component, perceivers with different degrees of teaching experience achieved comparable (medium) levels of accuracy in both studies, despite the different circumstances of the two studies. Moreover, perceived difficulty (study 1) and estimate of effort (study 2) were perceived with low accuracy. Both studies provide hints towards gender-specific distortions with respect to the level and differentiation component, but these effects were stronger for the pre-service teachers (study 2). Pre-service teachers furthermore overestimated the perceived task difficulty of the girls while at the same time underestimating the perceived task difficulty of the boys.

Given that in study 1, a research design was employed in which perceivers interact with previously unacquainted targets, this study can, according to the definition by Kenny & West (2008), also be referred to as a zero-acquaintance study. A comparison with study 2, in which experienced teachers indicate their judgments regarding students they are already familiar with through everyday classroom interaction, provides hints towards the role of acquaintance for teacher judgment accuracy. But an actual comparison was not possible given the differences in teaching experience between the perceivers in both studies. The present findings, however, further emphasize previous evidence showing that acquaintance only appears to play a minor role (Praetorius et al., 2015).

Taken together, the two studies constituting the first manuscript provide additional support for the subsistence of gender-related effects that can alter teacher judgments in the physics classroom. Manuscripts 2 and 3 further address the accuracy of perceiver judgments regarding different student characteristics based on brief videos (thin-slices of behavior) of previously unknown students (zero-acquaintance). These studies address potentially relevant relationship factors between judge and target, as well as perceptual effects related to students' sex, alongside the role of other cues in the perception of students' characteristics.

2.2 Manuscript 2: Teacher Judgments Based on Minimal information Using Social Accuracy Analyses

Recent research on teacher judgments has been applying alternative analytical strategies that go beyond the componential analyses suggested by Schrader (1989). Often, such approaches

involve complex multilevel models (e.g., Kolovou et al., 2021). Compared to earlier research, some of these models allow to investigate several student characteristics simultaneously (i.e., profile-based approaches), as well as to identify certain moderators that are expected to play a role for teacher judgment accuracy outcomes (see also Bhowmik et al., 2021; Südkamp et al., 2018). One example for such a mathematical model is the SAM (Biesanz, 2010, 2019), which forms the theoretical and methodological foundation of the second manuscript within this dissertation.

In previous teacher judgment accuracy research, the most commonly applied approach to determine a teacher's accuracy outcomes has been *variable-centered* or *trait-based*. In this analytical design, accuracy scores are obtained for each investigated characteristic separately using correlational analyses. However, the traditional correlational approach to teacher judgment accuracy has been criticized for the fact that only one variable or trait can be analyzed at a time (Karst, 2012; Südkamp & Praetorius, 2017) referring to Cronbach's (1955) classic componential analysis as the more comprehensive alternative (Cronbach's components of accuracy model). According to Cronbach, accuracies should not be based on global scores, i.e., solely considering the overall difference between judgments and actual characteristics, but rather be anchored in different accuracy components. He therefore introduced a more analytic treatment of accuracy for the case that several traits and several targets are judged simultaneously (i.e., *profile-based* approach or *person-centered* approach). In his variance analytic scheme, Cronbach (1955) therefore proposed four components, i.e., *elevation, differential elevation, stereotype accuracy* and *differential accuracy*. While elevation refers to perceivers' and targets' individual treatment of the rating scale itself, e.g., tendencies to favor the midpoint of a scale, differential elevation addresses perceivers' tendencies to make mild or strict judgments for certain targets compared to others. Stereotype accuracy can be obtained when a perceiver's subjective beliefs about a group of targets' characteristics, matches with those groups' actual profile of characteristics. Hence, the stereotype concept as it is utilized in this context is referred to as an individual's beliefs about certain groups, which, in the case of high stereotype accuracy, may respond to the groups' actual characteristics, (Ashmore & Del Boca, 1981; Jussim et al., 2015). Finally, differential accuracy describes the agreement between the judgment of an individuals' trait profile and the objective values of this individual's trait profile. Differential accuracy is considered the most relevant component because it measures the accuracy of judging targets with respect to specific traits after controlling for elevation, differential elevation and stereotype accuracy.

The SAM (Biesanz, 2010, 2019) is a newer model that successfully integrates the earlier models proposed by Cronbach (1955) and Kenny (Social Relations Model; 1988). SAM focuses on two components of Cronbach's model: differential accuracy now designated as *distinctive accuracy,* and stereotype accuracy now designated as *normative accuracy*. SAM uses a one-step analytical procedure which integrates both aspects of accuracy (Biesanz, 2010, 2019). In SAM, distinctive accuracy reflects an interaction between targets and traits, while differential elevation and stereotype accuracy represent main effects (see also Cronbach 1955). Distinctive accuracy may be influenced by normative accuracy when a perceiver's judgments of a target's individual trait profile reflects a common stereotype, i.e., a group's mean profile of traits (Jussim et al., 2009; Letzring, 2015). SAM can moreover support in identifying whether variables that are related to the perceiver, such as for instance teaching experience, and/or perceiver-target relationship variables, such as for instance liking, play a role for distinctive or normative accuracy. Such moderator variables that concern the judge-target relationship are referred to as *dyadic moderators*. Using SAM, one can also address questions that are related to the judgeability of different targets (i.e., *expressive accuracy*), that is, the accuracy with which a target is on average perceived by a set of perceivers. The SAM is computed via multilevel modeling, with the perceiver rating as the dependent and the judgment criterion as the independent variable. In this model, the specific perceiver-target dyad serves as the basic unit for the analysis (see manuscript 2 for a detailed description and mathematical illustration of the model). As a multilevel model, SAM accounts for dependencies in the data, for instance the fact that the group of students that is judged is most often part of a specific classroom that is taught by a specific teacher. Hence, one can assume that the students within each classroom are more similar than students across different classrooms. Retaining all information in one single analysis, SAM moreover increases statistical power compared to trait-wise approaches.

The design of this study, as well as the final study comprising the third manuscript (see 1.3), was inspired by the large existing body of research employing a thin-slice of behavior approach in personality and social psychology research, relying on brief target videos as stimulus material for perceiver judgments (Ambady & Rosenthal, 1992; Back & Nestler, 2016; Carney et al., 2007; Murphy et al., 2015). By employing brief video snippets of previously unknown students, the aim was to investigate essential aspects of first impression formation that translate to realistic classroom situations in which teachers come face to face with students they did not interact with

or meet before. Hence, the SAM was employed in this study to examine perceivers' first impressions regarding the academic self-concept, intrinsic motivation and intelligence of 10 students who served as targets. Based on the brief student videos, three groups of perceivers (pre-service teachers, experienced teachers and psychology students; $N = 285$) indicated their judgments based on brief 30-seconds long videos showing one student each while working on a physics experiment. In addition, perceivers indicated how much they liked each student in the video snippets and the perceived physical attractiveness for each student. In addition to the academic-self-concept and intrinsic motivation as motivational determinants of achievement, students' intelligence was added as a further criterion variable in this study as it marks a central prerequisite for learning that is implicitly evaluated by teachers on a regular basis. The shown impact of teacher expectations on students' self-perceptions moreover highlights the importance of this characteristic to be included in research on judgment accuracy (Machts et al., 2016; Pretzlik et al., 2003).

Whereas SAM analyses demonstrated high normative accuracy outcomes of the perceivers (i.e., knowledge concerning the average student's profile of characteristics), at the same time, distinctive accuracy (i.e., accurately perceiving individual students' unique trait profile) was low. Findings also showed that students who were likeable were perceived with lower distinctive accuracy. In addition, students that were evaluated as likeable and physically attractive students were perceived with higher normative accuracy and received more desirable evaluations by the perceivers. Perceiver-student personality similarity and perceivers' teaching experience did not play a significant role for either type of accuracy.

Aside from moderators concerning teacher and students separately, findings of this study suggest that it is particularly informative to include moderators concerning the teacher-student relationship when investigating teacher judgments. Taken together, these findings depict the potential of SAM to investigate such relationship moderators within one single analysis.

2.3 Manuscript 3: A Lens Model Approach to Teacher Judgment Accuracy

In this final study of the dissertation, it is suggested that the knowledge about the nature of interpersonal perception and judgment, obtained from the large body of research applying the Brunswikian theory and methods in social and personality psychology research (Cooksey, 1996; Karelaia & Hogarth, 2008; Kaufmann et al., 2013; Nestler & Back, 2013), can and should be transferred to the educational context to better understand teacher perceptual processes. Hence, the

lens model approach has a strong potential to support teachers in critically evaluating their judgmental tendencies, which, in turn, could be relevant for the process of achieving social justice in educational settings (see also a recent review by (Kaufmann, 2022).

BLM, as the central methodological representation of the Brunswikian theory, draws on the assumption that judgments regarding other individuals' characteristics that are not directly observable (e.g., personality traits, such as extraversion) are based on observable cues (e.g., behavior or physical appearance). Such cues are then utilized by the perceivers during the judgment process to draw conclusions about the criterion in question. In the educational context, an accurate judgment is therefore only possible when a) the environment contains relevant cues that are valid for the criterion variable (e.g., a student's values on a self-concept scale), also referred to as *cue validity* and b) the teacher is able to detect and utilize such valid cues in their judgment process, also referred to as *cue utilization*. The utilization of cues that are not valid for the specific criterion can indicate some form of perceptual bias or stereotype (Breil et al., 2021). For instance, teachers could utilize students' sex as a cue for intelligence or motivation, which would point towards a gender bias, as long as this tendency is not reflected in the actual target data (see also 1.1 for a more detailed elaboration of teacher expectancy effects, such as the gender bias). As a comprehensive conceptual model, BLM therefore represents a meaningful methodology that can explain high or low accuracy outcomes and provide hints for cognitive biases relevant to teacher judgment processes. Yet, BLM has rarely been employed in previous research on teacher judgments, irrespective of its potential to explain inter-individual differences in teacher judgment accuracy, some of which may be ascribed to differences in teachers' sensitivity for valid cues (Förster & Böhmer, 2017). The very few existing previous studies that applied BLM in teacher judgment research were synthesized by Kaufmann (2022) in a recent review, however, all available work has been focused on achievement judgments based on standardized tests. The research presented within this dissertation, in contrast, is mainly centered around teacher judgments of students' emotional and motivational characteristics as relevant prerequisites for learning and well-being in educational settings.

One central aim of this study was to identify relevant nonverbal cues expressed by students and the utilization of different cues by pre-service teachers and psychology students in their judgments of students' academic self-concept, motivation and intelligence. According to the thin-slice of behavior and zero-acquaintance approach outlined above (see 1.4), brief non-verbal video

clips of students served as stimulus material for the perceiver judgments and the students that served as targets in this investigation were unacquainted to the perceivers. This video material was subsequently coded by two independent raters, following the conception of a profound coding manual, which was inspired by previous work using BLM (Back & Nestler, 2016; Nestler et al., 2012). Despite students' spontaneous behavioral cues (e.g., friendly or attentive facial expressions or expressive gestures), students' physical cues (e.g., attractiveness, wearing eyeglasses, distinctive hair), which may influence teachers' judgments and the observability of students' individual characteristics, were investigated (see also Breil et al., 2021 for a recent example). Masculinity was included as cue in addition to students' biological sex to ensure that effects related to the social construction of gender can be identified. This was regarded as particularly relevant given that the physics classroom, in which the situations in the brief video clips took place, appears to be prone to gender-related expectations in favor of the male students, an effect which has also been evidenced by the research conducted within manuscript 1.

Findings of this study showed that while judgment accuracy was overall rather low, it varied substantially between constructs. Thus, highest accuracy values were found for the judgments of students' intrinsic motivation in physics (r = .23) and lowest for intelligence (r = -.03). Intrinsic motivation in physics was also the construct with the strongest predictability of the cues (R^2 = .61; r = .78) and cue-sensitivity (r = .37). These different levels of accuracy can be explained looking at the individual lens model parameters, such as cue validity, cue utilization, predictability, and consistency. Given that both predictability and cue sensitivity were highest for intrinsic motivation shows that (1) relevant cues were available in the environment and (2) perceivers were sufficiently sensitive to such valid cues. Overall, predictability for the set of cues was comparatively lower for students' intelligence and academic self-concept. At the same time, perceivers in parts gave importance to less valid cues, i.e., cues with a low ecological validity. For example, an attentive facial expression was utilized as indicator in the judgments of students' academic self-concept, whereas lens model analyses did not confirm the validity of this cue for this specific criterion. Instead, a friendly facial expression would have been a relevant cue. In other cases, perceivers used a valid cue, but for the judgment of the wrong criterion. For instance, perceivers expected students with a distinctive clothing style to be motivated for the subject of physics and to possess a strong general academic self-concept, but missed out on this cue where it actually would have been a valid indicator, namely students' domain-specific self-concept in physics. In other words, even though

predictability of the cues was comparatively low for this construct, perceivers relied on the cues nevertheless, and therefore cue sensitivity was overall also very low. Results further underline the importance of students' biological sex and masculinity for the judgment process. Finally, response consistencies (also referred to as the reliability of judgments) were high and comparative across all constructs, indicating that the perceivers in this study applied the same judgment strategy across all targets. Cases in which teachers do not utilize the same information in a similar fashion for each student could bear the risk for lower judgment accuracy outcomes and even a decrease in social justice, for instance when some students are preferred over other students (Kaufmann, 2022).

This final study within this dissertation illustrates the various possibilities of the lens model in creating an understanding of students' behavioral expression and teachers' perceptual processes. Such understanding appears to be of particular theoretical as well as practical value within the area of teacher judgment research, not least as it may inform initiatives aimed at the development of training programs to support teachers in reflecting upon their judgments and forming more accurate impressions (see also discussion for a more detailed elaboration of possible practical implications of the present research).

3 DISCUSSION

General Discussion

3.1 Summary

Teacher judgments and judgment accuracy have become a central issue for educational researchers over the past decades. The majority of previously conducted research on teacher judgment accuracy employed a variable-centered approach using correlations between teacher judgments of students' characteristics and their actual characteristics (rank order component). Such correlations have been documented not only for students' achievement and other cognitive characteristics, but also for students' self-concept, motivation and other non-cognitive characteristics. However, one question that has been largely unexplored is what information is used by teachers for their judgments and how valid or invalid this information is. In particular, there is no general agreement on factors influencing teacher judgments and judgment accuracy in a favorable or unfavorable way. Traditional correlational approaches limit the possibilities to include such moderating factors, particularly those concerning the perceiver-student dyad.

The overarching aim of the present dissertation was therefore to build on the existing research concerning teacher judgment accuracy, while at the same time extending the available knowledge about factors that can either foster or hinder accurate teacher judgments. Factors biasing or distorting judgments and thus decreasing accuracy, such as for instance the gender bias, were therefore of special interest in the present dissertation. To this end, theories and methodological approaches that have to date mostly been applied to interpersonal perception research in social and personality psychology and that have the potential to shed light on teacher judgment and judgment accuracy were adapted and implemented. These methods include the zero-acquaintance and thin-slices of behavior approach, as well as SAM und BLM.

Research conducted within this dissertation was focused on different aspects of the judgment process, with the aim to (1) identify possible gender-related cognitive distortions that play a role for teacher judgments in the physics classroom (manuscript 1), (2) investigate of the role of different (dyadic) moderators through applying SAM as a profile-based approach to teacher judgment accuracy (manuscript 2), and (3) reveal the cues that are used by teachers in their perceptual process and the validity of such cues for students' criterion variables applying lens model analyses. Table 1 illustrates a short summary of the findings which are elaborated in the three manuscripts that are comprising this dissertation.

Table 1

Short Summary of the Findings of the Present Dissertation

Manuscript	Aims	Method	Central findings
1	To investigate the role of students' sex for teacher judgments of students' emotional-motivational characteristics in the physics context	Pre-service (study 1; $N = 27$) and experienced teachers (study 2; $N = 207$) rated $N = 207$ students in respect to their academic self-concept, perceived task difficulty, and willingness to make an effort; application of all three accuracy components by Schrader & Helmke (1987) as well as *t*-tests to identify gender-related effects	Overall medium accuracy outcomes for pre-service teachers and experienced teachers, while students' academic self-concept was perceived most accurately; tendency to achieve higher accuracy in the ratings of boys (indication of a gender bias)
2	To examine teacher first impressions as well as to investigate the potential role of different moderating factors for teacher judgment accuracy applying SAM	Three groups of perceivers (student teachers, experienced teachers and psychology students; $N = 285$) rated students' ($N = 10$) profile of characteristics (academic self-concept, intrinsic motivation and intelligence) based on thin-slices of behavior; simultaneous investigation of distinctive and normative accuracy as well as judge and dyadic moderators (perceiver-target personality similarity, liking and teaching experience) using multilevel analyses	Perceivers were able to detect the average students' profile of characteristics (normative accuracy), but not students' unique personality profiles (distinctive accuracy); likeable and physically more attractive students were perceived with higher normative accuracy
3	To employ BLM to identify cues that teachers utilize in their first impressions of students and the validity of different cues for students' characteristics	Pre-service teachers and psychology students ($N = 102$) provided ratings of students' ($N = 45$) academic self-concept, intelligence and motivation based on brief nonverbal video clips; conduction of regression-based lens model analyses to obtain relevant lens model parameters such as cue validity, utilization, and sensitivity as well as predictability and consistency	Intrinsic motivation was perceived with the highest accuracy; perceivers utilized some valid cues but also failed to detect relevant information and/or allocated importance to non-valid cues in their judgments; high levels of consistency implies that perceivers applied the same judgment strategy across students and characteristics

3.2 Implications of the Present Dissertation

As outlined in the introduction part of this dissertation, teacher judgment accuracy can play a crucial role for students' emotional and motivational characteristics, as well as academic achievement and prospects. Recent research suggests that effects of teacher judgments on students' lives may even have been underestimated (see Bergold, 2023). The importance of accurate teacher judgments becomes particularly apparent in early teacher judgments, as those initial impressions can affect the subsequent teacher-student interaction, which, in case of erroneous judgments, can restrict learners' overall academic prospects and opportunities. The findings of the research comprising this dissertation expose the need to further investigate the rather complex cognitive processes involved in teacher judgment formation and give rise to several theoretical and practical implications.

The present findings are to some extent coherent with empirical evidence gained from previous and traditional teacher judgment accuracy research while at the same time, an interdisciplinary point of departure was chosen to look at teacher perceptual processes from a fresh angle, including aspects such as the disclosure of information teachers use for their judgments (BLM) as well as possible (dyadic) moderators that can play a role for teacher judgment accuracy utilizing statistically advanced procedures (SAM). The application of these two methods, combined with the zero-acquaintance and thin-slices of behavior approach provided further evidence for the role of cognitive distortions in respect to interpersonal judgments in educational settings, where faulty evaluations can have far-reaching consequences. Thus, both the findings of the application of the SAM model and the lens model study that comprise manuscripts two and three illustrate the role of students' physical appearance for teacher judgment accuracy. Beyond that, perceivers relied on students' sex as central cue for their judgments, an effect that was observed in all studies comprising manuscript one and three. Consequently, less accurate and thereby less favorable judgments were allocated to the female students in the sample. These effects were independent of the duration of acquaintance between perceivers and students, given that in the first investigations of this dissertation (manuscript 1), perceivers made their judgments after at least a 90-minute face-to-face interaction, compared to the judgments at zero-acquaintance in the lens model study (manuscript 3). The lens model parameter analysis moreover reflected that (1) some constructs seem to be more difficult to perceive accurately than others, (2) individual constructs are expressed in different sets of cues and (3) even in case valid cues are expressed by the students, it depends on

the teachers to pick up on those cues and correctly utilize them for their judgment. These aspects are coherent with the key aspects of Funder's RAM, suggesting that an accurate judgment is only possible if the perceiver manages to detect and correctly uses behaviors that are relevant to the criterion of interest and at the same time available to the perceiver during the observational process. Lens model analyses therefore aid in generating valuable insights regarding the information teachers use and integrate into their judgments, which in turn helps to identify the sources of individual teachers' inaccuracy (Kaufmann, 2022). Furthermore, as a profile-based approach, the SAM has shown to outperform traditional approaches in which teacher accuracy regarding one trait at a time is investigated. One particular strength of SAM is therefore the possibility to look at factors related to the specific teacher-student relationship and how this interaction can result in more or less accurate teacher judgments, while taking the dependencies in the data into account that need to be considered in teacher-student judgmental contexts. As common feature of all investigations that were carried out in the framework of this dissertation, the physics context turned out to be particularly suitable to reveal gender effects that can lead to inaccurate teacher judgments.

Not astonishingly, teachers' judgment accuracy is far from being perfect in the present research. Beyond furthering our theoretical understanding about teacher judgments and judgment formation processes, there are therefore also several practical implications of the present work that can aid in identifying possible means to support teachers in their judgment processes with the goal to arrive at more accurate (initial) judgments in the classroom. Considering the different angles, from which teacher judgment accuracy was investigated in this dissertation, the lens model appears to be particularly fruitful in respect to the construction of suitable training programs that help teachers to reduce cognitive biases and to rely on valid cues in their judgments. The findings presented in this dissertation therefore do not only underline the suitability of the lens model approach to further understand the information teachers utilize for their judgments in the classroom, particularly when meeting their students for the first time, it also offers connecting points for establishing and implementing specific training programs that enable teachers to arrive at more accurate judgments.

According to the lens model, an accurate perceiver uses valid cues (sensitivity) in a consistent manner (consistency) when making judgments about another person's characteristics. Lens model analyses in the present research, however, have shown that cue sensitivity was overall moderate to low, while at the same time, response consistency was rather high. This implies that

the perceivers in the present research were not sufficiently sensible to the cues that would have been relevant for the specific criterion. This perceptual tendency was consistent across the targets in the student sample and thus, appears to be a systemic concern. Therefore, a primary goal should be to sensitize teachers to the relevance of different cues for different student characteristics, an attempt that would require further in-depth research investigating the effectiveness of different possible approaches. One possibility could be the implementation of expert models as source of advice for teachers, which could be developed based on knowledge that has been generated through lens model analyses, however, one challenge being the acceptance of such expert models by the teachers (Kaufmann, 2022). In the context of her review, Kaufmann (2022) moreover proposes the application of a so-called individual participant data (IPD) meta-analytic approach. Over and above supporting teachers in identifying and selecting valid information for their judgments, creating awareness about the flexibility of expectations as well as the advantage of having positive expectations (rather than expecting too little) appears to be at least as relevant given the documented and potentially far-reaching consequences of inaccurate or negative teacher expectations (de Boer et al., 2010; Timmermans et al., 2021).

One central building block for the quests described above concerns the facilitation of teachers' engagement in reflectional processes, as it has been illustrated in Schön's (1983, 1987) perspective of the teacher as *reflective practitioner*. According to Schön, reflective practice incorporates thoughtful consideration of one's own experiences in applying knowledge to practice while being accompanied by professionals (Schön 1983, 1987). A training framework that would support teachers in learning from their experiences in different judgment situations could be an important step into this direction. To achieve this, it is crucial that teachers monitor their regular diagnostic activity and skills and consider these a central aspect of their professional role (Helmke et al., 2004).

A closely related aspect to consider in the construction of such teacher programs is the potential of feedback to induce knowledge updating processes and foster cue sensitivity (e.g., Hoffrage et al., 2000; Nestler et al., 2012). Feedback about the actual criterion values has shown to lead to knowledge updating processes, which also has been investigated in the context of research on the so-called *hindsight effect* (Hoffrage et al., 2000; Nestler et al., 2012; Pohl et al., 2003; Pohl, 2007). The hindsight effect describes a phenomenon where individuals are first asked about their judgments regarding a specific situation or target criterion and are subsequently given

feedback about the actual outcome. Afterwards, perceivers in those experiments are normally asked to recall their original judgments, which often reveals a shift in the judgments towards the actual outcome. Hence, perceivers tend to assume that they have known the actual result or been right about it from the very beginning. Given this observed shift towards more accurate assessments subsequent to receiving feedback the hindsight effect has also been described as a side-effect of successful learning (Nestler et al., 2012). Seminars or automated environments that give feedback to the perceivers upon their judgment could therefore aid in raising awareness among teachers about their own perceptional tendencies and habits. An example for an automated environment is the *simulated classroom* (Kaiser et al., 2012; Südkamp et al., 2008), a computer simulation of a classroom situation, in which the participating teachers or teacher students interact with simulated students before assessing their performance, which is experimentally manipulated by the program. Tools, such as the simulated classroom, could also aid in tracing and evaluating teachers' learning processes towards more accurate judgments. Furthermore, data can be gathered, such as for instance the information that teachers utilize in their judgments, that can support researchers in further understanding processes involved in teacher judgment formation. In sum, implementing feedback and reflective processes in the construction of teacher training programs designed to support teachers' perceptual accuracy appears to be central and relevant.

3.3 Limitations and Directions for Future Research

In the present dissertation, I have illustrated the possibilities of integrating the theoretical traditions related to person perception research in psychological science with research on teacher judgment accuracy, which normally is confined to the area of empirical educational science. However, despite the considerable potential that is paralleled by such an interdisciplinary approach, there are some limitations of the present work which need to be taken into consideration[2].

BLM assumes that perceivers apply a linear decision rule, with which they connect the individual cues in their judgment processes (Karelaia & Hogarth, 2008). However, whereas this assumption has typically shown to result in a good prediction of the judgments, it remains unclear if the reality of the actual cognitive processes is in fact linear given that to date, scientific evidence regarding the nature of the involved cognitive processes is very limited. In a nutshell, linear models,

[2] Limitations that concern methodological aspects in the individual studies, such as for instance sample size, have already been discussed in the individual manuscripts and are therefore not repeated within this section.

such as BLM, tend to offer a good approximation, even though the actual relationships may not be linear. In their meta-analysis, Karelaia and Hogarth (2008) therefore argue that whereas the vast body of existing lens model research has shown that linear models are indeed capable of providing solid representations of human judgment, there can be situations in which human decisions and judgment may be better described by non-linear processes (e.g., judgments under stress or time pressure). It is therefore important to bear in mind that BLM can and should be regarded as an approximation to actual and complex judgmental processes and that actual cognitive processes involved in interpersonal perception can to date not fully be comprehended. It is recommended that further research focuses on disentangling such complex cognitive processes.

Moreover, the application of the thin-slices of behavior and zero-acquaintance approach implies that there is a shift from knowledge- and experience-based judgments to judgments that resemble first impressions. Such first impressions can only be based on information that is available and present in the observational situation, that is, students' physical appearance and behavior. A number of studies have found that first impressions are formed rapidly and tend to inform subsequent expectations and judgments, which in turn can alter perceiver-target interactions and, for the case of teacher judgments, influence students' academic as well as socio-psychological and behavioral outcomes (e.g., Ambady & Rosenthal, 1992; Carney et al., 2007; Clifford & Walster, 1973; Darley & Fazio, 1980; de Boer et al., 2010; Harris & Garris, 2008; Human et al., 2013; Jussim, 1989; Murphy et al., 2014; Nickerson, 1998; Rubie-Davies, 2010; Timmermans et al., 2021; Wang et al., 2018). At the same time, there is still considerable uncertainty with regard to continuous judgmental processes. Hence, it would be interesting to investigate to what extent these first impressions steer and influence the processing of subsequent information, the extent to which they are stable or get revised and, related to this, how long the effectiveness of teachers' first impressions last. The little available research addressing those questions has been focused on the stability of teacher expectations (e.g., Kuklinski & Weinstein, 2000; Timmermans et al., 2021; Timmermans & Rubie-Davies, 2018; Wang et al., 2020), however, evidence remains inconsistent and teachers continued expectancy effects were mainly assessed in the context of students' performance and academic achievement. There is furthermore room for further progress in determining possible differences between teachers in respect to the processes outlined above and how this difference can be explained by other factors.

A teacher's inaccuracy can take different forms, such as overly positive or negative evaluations. A more recent wave of research has therefore included possible effects of teacher under- and overestimations as distinct manifestations of accuracy, indicating that an overly positive impression may be more beneficial for students than if a student gets underestimated (Bergold, 2023; Förster et al., 2022; Rubie-Davies et al., 2006; Urhahne et al., 2011). Future research on (early) teacher impressions should take this extended understanding of teacher judgment accuracy into account and further investigate contexts that lead to over- and underestimations of students' academic ability and emotional or motivational characteristics.

Taken together, these questions represent a vital issue for future research, ideally applying a longitudinal research design in which teachers' first impressions can be set in relation to subsequent expectations and judgments, teacher-student interactions, students' academic achievement and behavioral as well as emotional and motivational development over time.

3.4 Conclusion

The present dissertation highlights the importance of taking a look at the larger picture of teacher judgment accuracy by combining different areas of psychological and educational science, while it has been shown to be fruitful to move beyond the traditional procedures to address teacher judgment accuracy through the exploration of different methodological possibilities. By this, questions regarding teachers' perceptual processes can be addressed from different methodological perspectives, and, moreover, individual outcomes can be validated. Approaches addressing the accuracy of personality judgments and interpersonal perception processes established in social and personality psychology have demonstrated a considerable potential to further our understanding of teacher judgments and teacher judgment accuracy.

An application of a zero-acquaintance and thin-slices of behavior approach ensures that teachers possess no prior knowledge about the individual students when making their judgments. Teacher judgments at zero-acquaintance therefore resemble situations in which a teacher is meeting a student for the first time. Such early impressions can steer a teacher's perception and assimilation of information, and thus be a starting point for forming knowledge-based judgments which are an essential component of teachers' expertise and experience. Furthermore, first impressions and early judgments may influence subsequent teacher-student interactions, and by this, potentially also learners' long-time academic perspectives. Through the application of this method, particular

attention can be allocated to the judgment process itself, independent of possible influences based on previous teacher-student interaction. Hence, focusing on spontaneous teacher judgments that qualify as first impressions entails the possibility to understand the genesis of teacher judgments. In combination with the investigation of individual and relationship moderators and relevant cognitive biases, this procedure can then aid in generating a profound comprehension of (early) teacher judgment processes.

As central examples for methodological approaches that have been established in social and personality psychology, the application of SAM and BLM can contribute to expand our knowledge and understanding of teachers' diagnostic competence. On the one hand, BLM disentangles the conventional accuracy measure (rank order component) in components of the environment (ecological validity) and the perceiver (response consistency and sensitivity). On the other hand, SAM resembles a profile-based approach that introduces alternative and more comprehensive concepts of accuracy (i.e., normative and distinctive accuracy) that are not least beneficial for statistical reasons. The application of BLM can furthermore support in identifying the type of information that is used by teachers during the judgment process, which is particularly important when dealing with thin-slice or zero-acquaintance judgments and nonverbal stimulus material such as student expressive behavior in brief video clips.

Given the possibly far-reaching consequences of inaccurate teacher judgments, paving the way to arrive at more accurate judgments should be of a central concern. It is hoped that insights gained through this research will (1) be used as a starting point for further integrative research on teacher judgment accuracy and processes utilizing SAM or BLM and (2) be of value for the development of teacher education and professional development programs devoted to fostering teachers' reflective practices towards their own perceptual processes.

4 REFERENCES

References

Albright, L., Kenny, D., & Malloy, T. (1988). Consensus in Personality Judgments at Zero Acquaintance. *Journal of Personality and Social Psychology*, *55*, 387–395. https://doi.org/10.1037/0022-3514.55.3.387

Allport, G. W. (1937). *Personality: A psychological interpretation* (pp. xiv, 588). Holt.

Ambady, N., Bernieri, F. J., & Richeson, J. A. (2000). Toward a histology of social behavior: Judgmental accuracy from thin slices of the behavioral stream. In M. P. Zanna, *Advances in Experimental Social Psychology* (Vol. 32, pp. 201–271). Academic Press. https://doi.org/10.1016/S0065-2601(00)80006-4

Ambady, N., Hallahan, M., & Rosenthal, R. (1995). On judging and being judged accurately in zero-acquaintance situations. *Journal of Personality and Social Psychology*, *69*(3), 518–529. https://doi.org/10.1037/0022-3514.69.3.518

Ambady, N., & Rosenthal, R. (1992). Thin slices of expressive behavior as predictors of interpersonal consequences: A meta-analysis. *Psychological Bulletin*, *111*(2), 256–274. https://doi.org/10.1037/0033-2909.111.2.256

Ambady, N., & Rosenthal, R. (1993). Half a minute: Predicting teacher evaluations from thin slices of nonverbal behavior and physical attractiveness. *Journal of Personality and Social Psychology*, *64*(3), 431–441.

Anderson, N. H., & Barrios, A. A. (1961). Primacy effects in personality impression formation. *The Journal of Abnormal and Social Psychology*, *63*(2), 346–350. https://doi.org/10.1037/h0021966

Artelt, C. (2015). Teacher Judgments and their Role in the Educational Process. In *Emerging Trends in the Social and Behavioral Sciences: An interdisciplinary, searchable, and linkable resource* (pp. 1–16). John Wiley & Sons. https://doi.org/10.1002/9781118900772.etrds0402

Asch, S. E. (1946). Forming impressions of personality. *The Journal of Abnormal and Social Psychology*, *41*(3), 258–290. https://doi.org/10.1037/h0055756

Ashmore, R. D., & Del Boca, F. K. (1981). Conceptual approaches to stereotypes and stereotyping. In D. L. Hamilton (Ed.), *Cognitive processes in stereotyping and intergroup behavior* (pp. 1–31). Lawrence Erlbaum Associates Publishers.

Babad, E., Bernieri, F., & Rosenthal, R. (1989a). Nonverbal communication and leakage in the behavior of biased and unbiased teachers. *Journal of Personality and Social Psychology*, *56*(1), 89–94. https://doi.org/10.1037/0022-3514.56.1.89

Babad, E., Bernieri, F., & Rosenthal, R. (1989b). When Less Information Is More Informative: Diagnosing Teacher Expectations from Brief Samples of Behaviour. *British Journal of Educational Psychology*, *59*(3), 281–295. https://doi.org/10.1111/j.2044-8279.1989.tb03103.x

Back, M. D., & Nestler, S. (2016). Accuracy of judging personality. In J. A. Hall, M. Schmid Mast, & T. V. West (Eds.), *The Social Psychology of Perceiving Others Accurately* (pp. 98–125). Cambridge University Press.

Baumert, P. D. J., & Kunter, D. M. (2006). Stichwort: Professionelle Kompetenz von Lehrkräften [Keyword: Professional Competence of Teachers]. *Zeitschrift für Erziehungswissenschaft*, *9*(4), 469–520. https://doi.org/10.1007/s11618-006-0165-2

Beer, A., & Watson, D. (2010). The effects of information and exposure on self-other agreement. *Journal of Research in Personality*, *44*(1), 38–45. https://doi.org/10.1016/j.jrp.2009.10.002

Begeny, J. C., & Buchanan, H. (2010). Teachers' judgments of students' early literacy skills measured by the Early Literacy Skills Assessment: Comparisons of teachers with and without assessment administration experience. *Psychology in the Schools*, *47*(8), 859–868. https://doi.org/10.1002/pits.20509

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*(1), 43–55. https://doi.org/10.1037/1045-3830.23.1.43

Begeny, J. C., Krouse, H. E., Brown, K. G., & Mann, C. M. (2011). Teacher Judgments of Students' Reading Abilities Across a Continuum of Rating Methods and Achievement Measures. *School Psychology Review*, *40*(1), 23–38. https://doi.org/10.1080/02796015.2011.12087726

Begrich, L., Kuger, S., Klieme, E., & Kunter, M. (2021). At a first glance – How reliable and valid is the thin slices technique to assess instructional quality? *Learning and Instruction*, *74*, 101466. https://doi.org/10.1016/j.learninstruc.2021.101466

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

Bergold, S., R. (2023). Teacher judgments predict developments in adolescents' school performance, motivation, and life satisfaction. *Journal of Educational Psychology*, *115*(4), 642–664. https://doi.org/10.1037/edu0000786

Bhowmik, C. V., Nestler, S., Schrader, F.-W., Praetorius, A.-K., Biesanz, J. C., & Back, M. D. (2021). Teacher judgments at zero-acquaintance: A social accuracy analysis. *Contemporary Educational Psychology*, *65*, Article 101965. https://doi.org/10.1016/j.cedpsych.2021.101965

Biesanz, J. C. (2010). The social accuracy model of interpersonal perception: Assessing individual differences in perceptive and expressive accuracy. *Multivariate Behavioral Research*, *45*(5), 853–885. https://doi.org/10.1080/00273171.2010.519262

Biesanz, J. C. (2019). The Social Accuracy Model. In T. D. Letzring & J. S. Spain (Eds.), *The Oxford Handbook of Accurate Personality Judgment*. Oxford University Press. https://doi.org/10.1093/oxfordhb/9780190912529.013.5

Biesanz, J. C., West, S. G., & Millevoi, A. (2007). What do you learn about someone over time? The relationship between length of acquaintance and consensus and self-other agreement in judgments of personality. *Journal of Personality and Social Psychology*, *92*(1), 119–135. https://doi.org/10.1037/0022-3514.92.1.119

Blackman, M. C., & Funder, D. C. (1998). The effect of information on consensus and accuracy in personality judgment. *Journal of Experimental Social Psychology*, *34*(2), 164–181.

Boer, H. de, Bosker, R. J., & Werf, M. P. van der. (2010). Sustainability of Teacher Expectation Bias Effects on Long-term Student Performance. *Journal of Educational Psychology*, *102*(1), 168–179.

Bonefeld, M., Dickhäuser, O., & Karst, K. (2020). Do preservice teachers' judgments and judgment accuracy depend on students' characteristics? The effect of gender and immigration background. *Social Psychology of Education*, *23*(1), 189–216. https://doi.org/10.1007/s11218-019-09533-2

Borkenau, P., Mauer, N., Riemann, R., Spinath, F. M., & Angleitner, A. (2004). Thin Slices of Behavior as Cues of Personality and Intelligence. *Journal of Personality and Social Psychology*, *86*(4), 599–614. https://doi.org/10.1037/0022-3514.86.4.599

Breil, S. M., Osterholz, S., Nestler, S., & Back, M. D. (2021). Contributions of Nonverbal Cues to the Accurate Judgment of Personality Traits. In T. D. Letzring & J. S. Spain (Eds.), *The Oxford Handbook of Accurate Personality Judgment* (pp. 195–218). Oxford University Press. https://doi.org/10.1093/oxfordhb/9780190912529.013.13

Brophy, J. E. (1983). Research on the self-fulfilling prophecy and teacher expectations. *Journal of Educational Psychology*, *75*(5), 631–661. https://doi.org/10.1037/0022-0663.75.5.631

Brunswik, E. (1956). *Perception and the Representative Design of Psychological Experiments*. University of California Press.

Carney, D. R., Colvin, C. R., & Hall, J. A. (2007). A thin slice perspective on the accuracy of first impressions. *Journal of Research in Personality*, *41*(5), 1054–1072. https://doi.org/10.1016/j.jrp.2007.01.004

Cate, I. P., Krolak-Schwerdt, S., Glock, S., & Markova, M. (2014). Improving Teachers' Judgments: Obtaining Change Through Cognitive Processes. In *Teachers' Professional Development* (pp. 45–61). Brill. https://brill.com/display/book/9789462095366/BP000005.xml

Chaplin, W. F., & Panter, A. T. (1993). Shared Meaning and the Convergence among Observers' Personality Descriptions. *Journal of Personality*, *61*(4), 553–585. https://doi.org/10.1111/j.1467-6494.1993.tb00782.x

Clifford, M. M., & Walster, E. (1973). The Effect of Physical Attractiveness on Teacher Expectations. *Sociology of Education*, *46*(2), 248–258. https://doi.org/10.2307/2112099

Colvin, C. R. (1993). " Judgable" people: Personality, behavior, and competing explanations. *Journal of Personality and Social Psychology*, *64*(5), 861–873. https://doi.org/10.1037/0022-3514.64.5.861

Connelly, B. S., & Ones, D. S. (2010). An other perspective on personality: Meta-analytic integration of observers' accuracy and predictive validity. *Psychological Bulletin*, *136*(6), 1092–1122. https://doi.org/10.1037/a0021212

Cooksey, R. W. (1996). The Methodology of Social Judgement Theory. *Thinking & Reasoning*, *2*(2–3), 141–174. https://doi.org/10.1080/135467896394483

Corno, L. (2008). On teaching adaptively. *Educational Psychologist*, *43*(3), 161–173. https://doi.org/10.1080/00461520802178466

Cronbach, L. (1955). Processes affecting scores on" understanding of others" and" assumed similarity.". *Psychological Bulletin*, *52*(3), 177–193. https://doi.org/10.1037/h0044919

Dapp, L. C., & Roebers, C. M. (2021). Metacognition and self-concept: Elaborating on a construct relation in first-grade children. *PLOS ONE*, *16*(4), e0250845. https://doi.org/10.1371/journal.pone.0250845

Darley, J. M., & Fazio, R. H. (1980). Expectancy confirmation processes arising in the social interaction sequence. *American Psychologist*, *35*(10), 867–881. https://doi.org/10.1037/0003-066X.35.10.867

de Boer, H., Bosker, R. J., & van der Werf, M. P. C. (2010). Sustainability of teacher expectation bias effects on long-term student performance. *Journal of Educational Psychology*, *102*(1), 168–179. https://doi.org/10.1037/a0017289

DePaulo, B. M. (1992). Nonverbal behavior and self-presentation. *Psychological Bulletin*, *111*(2), 203–243. https://doi.org/10.1037/0033-2909.111.2.203

Doherty, J., & Conolly, M. (1985). How accurately can primary school teachers predict the scores of their pupils in standardised tests of attainment? A study of some non-cognitive factors that influence specific judgements. *Educational Studies*, *11*(1), 41–60. https://doi.org/10.1080/0305569850110105

Doherty, M. E., & Kurz, E. M. (1996). Social Judgement Theory. *Thinking & Reasoning*, *2*(2–3), 109–140. https://doi.org/10.1080/135467896394474

Doornkamp, L., Van der Pol, L. D., Groeneveld, S., Mesman, J., Endendijk, J. J., & Groeneveld, M. G. (2022). Understanding gender bias in teachers' grading: The role of gender stereotypical beliefs. *Teaching and Teacher Education*, *118*, 103826. https://doi.org/10.1016/j.tate.2022.103826

Dovidio, J. F., Piliavin, J. A., Schroeder, D. A., & Penner, L. (2006). *The social psychology of prosocial behavior*. Lawrence Erlbaum Associates Publishers.

Dusek, J. B., & Joseph, G. (1983). The bases of teacher expectancies: A meta-analysis. *Journal of Educational Psychology*, *75*(3), 327–346. https://doi.org/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*(3), 247–265. https://doi.org/10.1002/pits.20147

Efklides, A. (2006). Metacognitive Experiences: The Missing Link in the Self-Regulated Learning Process. *Educational Psychology Review*, *18*(3), 287–291. https://doi.org/10.1007/s10648-006-9021-4

Efklides, A. (2011). Interactions of Metacognition With Motivation and Affect in Self-Regulated Learning: The MASRL Model. *Educational Psychologist*, *46*(1), 6–25. https://doi.org/10.1080/00461520.2011.538645

Efklides, A. (2017). Affect, Epistemic Emotions, Metacognition, and Self-Regulated Learning. *Teachers College Record*, *119*(13), 1–22. https://doi.org/10.1177/016146811711901302

Efklides, A., & Tsiora, A. (2002). Metacognitive experiences, self-concept, and self-regulation. *Psychologia: An International Journal of Psychology in the Orient*, *45*(4), 222–236. https://doi.org/10.2117/psysoc.2002.222

Feingold, A. (1992). Good-looking people are not what we think. *Psychological Bulletin*, *111*(2), 304–341. https://doi.org/10.1037/0033-2909.111.2.304

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*(1), 527–561. https://doi.org/10.1006/obhd.2001.2981

Flavell, J. H. (1979). Metacognition and cognitive monitoring: A new area of cognitive–developmental inquiry. *American Psychologist*, *34*(10), 906–911. https://doi.org/10.1037/0003-066X.34.10.906

Förster, N., & Böhmer, I. (2017). Das Linsenmodell - Grundlagen und exemplarische Anwendungen in der pädagogisch-psychologischen Diagnostik [The lens model – principals and exemplary applications in peadagogical-psychological diagnostics]. In A. Südkamp & A.-K. Praetorius (Eds.), *Diagnostische Kompetenz von Lehrkräften. Theoretische und methodische Weiterentwicklungen* (pp. 46–50). Waxmann.

Förster, N., Humberg, S., Hebbecker, K., Back, M. D., & Souvignier, E. (2022). Should teachers be accurate or (overly) positive? A competitive test of teacher judgment effects on students' reading progress. *Learning and Instruction*, *77*, 101519. https://doi.org/10.1016/j.learninstruc.2021.101519

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

Fultz, A. A., Stosic, M. D., & Bernieri, F. J. (2023). Nonverbal Expressivity, Physical Attractiveness, and Liking: First Impression to Established Relationship. *Journal of Nonverbal Behavior*. https://doi.org/10.1007/s10919-023-00444-7

Funder, D. C. (1995). On the accuracy of personality judgment: A realistic approach. *Psychological Review*, *102*(4), 652–670. https://doi.org/10.1037/0033-295X.102.4.652

Funder, D. C., & Colvin, C. R. (1991). Explorations in behavioral consistency: Properties of persons, situations, and behaviors. *Journal of Personality and Social Psychology*, *60*(5), 773–794. https://doi.org/10.1037/0022-3514.60.5.773

Funder, D. C., & Dobroth, K. M. (1987). Differences between traits: Properties associated with interjudge agreement. *Journal of Personality and Social Psychology*, *52*(2), 409–418. https://doi.org/10.1037/0022-3514.52.2.409

Funder, D. C., Kolar, D. W., & Blackman, M. C. (1995). Agreement among judges of personality: Interpersonal relations, similarity, and acquaintanceship. *Journal of Personality and Social Psychology*, *69*(4), 656–672. https://doi.org/10.1037/0022-3514.69.4.656

Garcia, E. B., Sulik, M. J., & Obradović, J. (2019). Teachers' perceptions of students' executive functions: Disparities by gender, ethnicity, and ELL status. *Journal of Educational Psychology*, *111*, 918–931. https://doi.org/10.1037/edu0000308

Geukes, K., Nestler, S., Hutteman, R., Küfner, A. C. P., & Back, M. D. (2017). Trait personality and state variability: Predicting individual differences in within- and cross-context fluctuations in affect, self-evaluations, and behavior in everyday life. *Journal of Research in Personality*, *69*, 124–138. https://doi.org/10.1016/j.jrp.2016.06.003

Glock, S. (2016). Does ethnicity matter? The impact of stereotypical expectations on in-service teachers' judgments of students. *Social Psychology of Education*, *19*(3), 493–509. https://doi.org/10.1007/s11218-016-9349-7

Glock, S., Krolak-Schwerdt, S., Klapproth, F., & Böhmer, M. (2013). Beyond judgment bias: How students' ethnicity and academic profile consistency influence teachers' tracking

judgments. *Social Psychology of Education*, *16*(4), 555–573. https://doi.org/10.1007/s11218-013-9227-5

Graves, A. L., Hoshino-Browne, E., & Lui, K. P. H. (2017). Swimming against the Tide: Gender Bias in the Physics Classroom. *Journal of Women and Minorities in Science and Engineering*, *23*(1). https://doi.org/10.1615/JWomenMinorScienEng.2017013584

Hammond, K. R., Stewart, T. R., Brehmer, B., & Steinmann, D. O. (1986). Social Judgment Theory. In H. R. Arkes & K. R. Hammond (Eds.), *Judgment and decision making: An interdisciplinary reader* (pp. 56–76). Cambridge University Press.

Hand, S., Rice, L., & Greenlee, E. (2017). Exploring teachers' and students' gender role bias and students' confidence in STEM fields. *Social Psychology of Education*, *20*(4), 929–945. https://doi.org/10.1007/s11218-017-9408-8

Hardy, I., Decristan, J., & Klieme, E. (2019). Adaptive teaching in research on learning and instruction. *Journal for Educational Research Online*, *11*(2), 169–191. https://doi.org/10.25656/01:18004

Harris, M. J., & Garris, C. P. (2008). You never get a second chance to make a first impression: Behavioral consequences of first impressions. In N. Ambady & J. J. Skowronski (Eds.), *First impressions* (pp. 147–168). Guilford Publications.

Hattie, J. (2009). *Visible Learning: A Synthesis of Over 800 Meta-Analyses Relating to Achievement* (1st ed.). Routledge.

Hattie, J. (2012). *Visible Learning for Teachers: Maximizing Impact on Learning*. Routledge.

Heller, K. A., Finsterwald, M., & Ziegler, A. (2001). Implicit theories of German mathematics and physics teachers on gender specific giftedness and motivation. *Psychologische Beiträge*, *43*(1), 172–189.

Helmke, A., Hosenfeld, I., & Schrader, F. (2004). Vergleichsarbeiten als Instrument zur Verbesserung der Diagnosekompetenz von Lehrkräften [Comparative tests as instrument to improve teachers' diagnostic competence]. In R. Arnold & C. Griese (Eds.), *Schulleitung und Schulentwicklung* (pp. 119–144). Schneider-Verlag.

Helmke, A., & Schrader, F.-W. (1987). Interactional effects of instructional quality and teacher judgement accuracy on achievement. *Teaching and Teacher Education*, *3*(2), 91–98.

Herppich, S., Praetorius, A.-K., Förster, N., Glogger-Frey, I., Karst, K., Leutner, D., Behrmann, L., Böhmer, M., Ufer, S., Klug, J., Hetmanek, A., Ohle, A., Böhmer, I., Karing, C.,

Kaiser, J., & Südkamp, A. (2018). Teachers' assessment competence: Integrating knowledge-, process-, and product-oriented approaches into a competence-oriented conceptual model. *Teaching and Teacher Education*, *76*, 181–193. https://doi.org/10.1016/j.tate.2017.12.001

Herwartz-Emden, L., Schurt, V., & Waburg, W. (2012). *Mädchen und Jungen in Schule und Unterricht* [Girls and boys in school and during instruction]. Kohlhammer Verlag.

Hirschmüller, S., Egloff, B., Nestler, S., & Back, M. D. (2013). The dual lens model: A comprehensive framework for understanding self–other agreement of personality judgments at zero acquaintance. *Journal of Personality and Social Psychology*, *104*(2), 335–353. https://doi.org/10.1037/a0030383

Hofer, S. I. (2015). Studying Gender Bias in Physics Grading: The role of teaching experience and country. *International Journal of Science Education*, *37*(17), 2879–2905. https://doi.org/10.1080/09500693.2015.1114190

Hoffrage, U., Hertwig, R., & Gigerenzer, G. (2000). Hindsight bias: A by-product of knowledge updating? *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *26*(3), 566–581. https://doi.org/10.1037//0278-7393.26.3.566

Hoge, R. D., & Coladarci, T. (1989). Teacher-Based Judgments of Academic Achievement: A Review of Literature. *Review of Educational Research*, *59*(3), 297–313. https://doi.org/10.2307/1170184

Holder, K., & Kessels, U. (2017). Gender and ethnic stereotypes in student teachers' judgments: A new look from a shifting standards perspective. *Social Psychology of Education*, *20*(3), 471–490. https://doi.org/10.1007/s11218-017-9384-z

Hornstra, L., Stroet, K., van Eijden, E., Goudsblom, J., & Roskamp, C. (2018). Teacher expectation effects on need-supportive teaching, student motivation, and engagement: A self-determination perspective. *Educational Research and Evaluation*, *24*(3–5), 324–345. https://doi.org/10.1080/13803611.2018.1550841

Huang, C. (2011). Self-concept and academic achievement: A meta-analysis of longitudinal relations. *Journal of School Psychology*, *49*(5), 505–528. https://doi.org/10.1016/j.jsp.2011.07.001

Human, L. J., & Biesanz, J. C. (2011). Through the looking glass clearly: Accuracy and assumed similarity in well-adjusted individuals' first impressions. *Journal of Personality and Social Psychology*, *100*(2), 349–364. https://doi.org/10.1037/a0021850

Human, L. J., & Biesanz, J. C. (2013). Targeting the good target: An integrative review of the characteristics and consequences of being accurately perceived. *Personality and Social Psychology Review*, *17*(3), 248–272. https://doi.org/10.1177/1088868313495593

Human, L. J., Sandstrom, G. M., Biesanz, J. C., & Dunn, E. W. (2013). Accurate First Impressions Leave a Lasting Impression: The Long-Term Effects of Distinctive Self-Other Agreement on Relationship Development. *Social Psychological and Personality Science*, *4*(4), 395–402. https://doi.org/10.1177/1948550612463735

Ingenkamp, K. (1972). *Die Fragwürdigkeit der Zensurengebung: Texte und Untersuchungsberichte* [The questionability of grades: Texts and investigation reports]. Beltz.

Jansen, M., Schroeders, U., Lüdtke, O., & Marsh, H. W. (2019). The dimensional structure of students' self-concept and interest in science depends on course composition. *Learning and Instruction*, *60*, 20–28. https://doi.org/10.1016/j.learninstruc.2018.11.001

Jiang, S., Paxton, A., Ramírez-Esparza, N., & García-Sierra, A. (2023). Toward a dynamic approach of person perception at zero acquaintance: Applying recurrence quantification analysis to thin slices. *Acta Psychologica*, 103866. https://doi.org/10.1016/j.actpsy.2023.103866

John, O., Naumann, L., & Soto, C. (2008). Paradigm shift to the integrative big five trait taxonomy: History, measurement, and conceptual issues. In *Handbook of personality: Theory and research* (3rd ed., pp. 114–158). The Guliford Press.

Jussim, L. (1989). Teacher Expectations: Self-Fulfilling Prophecies, Perceptual Biases, and Accuracy. *Journal of Personality and Social Psychology*, *57*(3), 469–480.

Jussim, L., Cain, T. R., Crawford, J. T., Harber, K., & Cohen, F. (2009). The unbearable accuracy of stereotypes. In T. D. Nelson (Ed.), *Handbook of prejudice, stereotyping, and discrimination* (pp. 199–227). Psychology Press. https://doi.org/10.4324/9781841697772

Jussim, L., Crawford, J. T., & Rubinstein, R. S. (2015). Stereotype (In)Accuracy in Perceptions of Groups and Individuals. *Current Directions in Psychological Science*, *24*(6), 490–497. https://doi.org/10.1177/0963721415605257

Jussim, L., & Eccles, J. S. (1992). Teacher expectations: II. Construction and reflection of student achievement. *Journal of Personality and Social Psychology*, *63*(6), 947–961. https://doi.org/10.1037/0022-3514.63.6.947

Jussim, L., & Harber, K. D. (2005). Teacher expectations and self-fulfilling prophecies: Knowns and unknowns, resolved and unresolved controversies. *Personality and Social Psychology Review: An Official Journal of the Society for Personality and Social Psychology, Inc*, *9*(2), 131–155. https://doi.org/10.1207/s15327957pspr0902_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 [About the relationship of intelligence and judgment accuracy in judgments of students' academic performance in the simulated classroom]. *Zeitschrift Für Pädagogische Psychologie*, *26*(4), 251–261. https://doi.org/10.1024/1010-0652/a000076

Karelaia, N., & Hogarth, R. M. (2008). Determinants of linear judgment: A meta-analysis of lens model studies. *Psychological Bulletin*, *134*(3), 404–426. https://doi.org/10.1037/0033-2909.134.3.404

Karst, K. (2012). *Kompetenzmodellierung des diagnostischen Urteils von Grundschullehrern*. Waxmann Verlag.

Kaufmann, E. (2022). Lens model studies: Revealing teachers' judgements for teacher education. *Journal of Education for Teaching*, *49*(2), 236–251. https://doi.org/10.1080/02607476.2022.2061336

Kaufmann, E. (2020). How accurately do teachers' judge students? Re-analysis of Hoge and Coladarci (1989) meta-analysis. *Contemporary Educational Psychology*, *63*, 101902. https://doi.org/10.1016/j.cedpsych.2020.101902

Kaufmann, E., Reips, U.-D., & Wittmann, W. W. (2013). A Critical Meta-Analysis of Lens Model Studies in Human Judgment and Decision-Making. *PLoS ONE*, *8*(12), e83528. https://doi.org/10.1371/journal.pone.0083528

Kenny, D. A., & West, T. V. (2008). Zero acquaintance: Definitions, statistical model, findings, and process. In *First impressions* (pp. 129–146). Guilford Publications. https://doi.org/10.1037/10559-014

Kessels, U., Warner, L. M., Holle, J., & Hannover, B. (2008). Identitätsbedrohung durch positives schulisches Leistungsfeedback [Identity threat through positive performance feedback in school]. *Zeitschrift Für Entwicklungspsychologie und Pädagogische Psychologie*, *40*(1), 22–31. https://doi.org/10.1026/0049-8637.40.1.22

Kilianski, S. E. (2008). Who do you think I think I am? Accuracy in perceptions of others' self-esteem. *Journal of Research in Personality*, *42*(2), 386–398. https://doi.org/10.1016/j.jrp.2007.07.004

Kolovou, D., Naumann, A., Hochweber, J., & Praetorius, A.-K. (2021). Content-specificity of teachers' judgment accuracy regarding students' academic achievement. *Teaching and Teacher Education*, *100*, 103298. https://doi.org/10.1016/j.tate.2021.103298

Kost-Smith, L. E., Pollock, S. J., & Finkelstein, N. D. (2010). Gender disparities in second-semester college physics: The incremental effects of a "smog of bias". *Physical Review Special Topics - Physics Education Research*, *6*(2), 020112. https://doi.org/10.1103/PhysRevSTPER.6.020112

Kriegbaum, K., Steinmayr, R., & Spinath, B. (2019). Longitudinal reciprocal effects between teachers' judgments of students' aptitude, students' motivation, and grades in math. *Contemporary Educational Psychology*, *59*, 101807. https://doi.org/10.1016/j.cedpsych.2019.101807

Kronig, W., & Ingenkamp, K. (2022). Wenn die Schulklasse die Note mitbestimmt [When students in the classroom influence the grade]. *Friedrich Jahresheft*, *2022*(1), 26–31.

Kuklinski, M., & Weinstein, R. (2000). Classroom and Grade Level Differences in the Stability of Teacher Expectations and Perceived Differential Teacher Treatment. *Learning Environments Research*, *3*(1), 1–34. https://doi.org/10.1023/A:1009904718353

Lansu, T. A. M., & van den Berg, Y. H. M. (2020). Thin-Slice Judgments of Children's Social Status and Behavior. *The Journal of Experimental Education*, *90*(4), 884–897. https://doi.org/10.1080/00220973.2020.1808943

Leaper, C., & Starr, C. R. (2019). Helping and Hindering Undergraduate Women's STEM Motivation: Experiences With STEM Encouragement, STEM-Related Gender Bias, and Sexual Harassment. *Psychology of Women Quarterly*, *43*(2), 165–183. https://doi.org/10.1177/0361684318806302

Leslie, S.-J., Cimpian, A., Meyer, M., & Freeland, E. (2015). Expectations of brilliance underlie gender distributions across academic disciplines. *Science*, *347*(6219), 262–265. https://doi.org/10.1126/science.1261375

Letzring, T. D. (2008). The good judge of personality: Characteristics, behaviors, and observer accuracy. *Journal of Research in Personality*, *42*(4), 914–932. https://doi.org/10.1016/j.jrp.2007.12.003

Letzring, T. D. (2010). The effects of judge-target gender and ethnicity similarity on the accuracy of personality judgments. *Social Psychology*, *41*(1), 42–51. https://doi.org/10.1027/1864-9335/a000007

Letzring, T. D. (2015). Observer judgmental accuracy of personality: Benefits related to being a good (normative) judge. *Journal of Research in Personality*, *54*, 51–60. https://doi.org/10.1016/j.jrp.2014.05.001

Letzring, T. D., Murphy, N. A., Allik, J., Beer, A., Zimmermann, J., & Leising, D. (2021). The Judgment of Personality: An Overview of Current Empirical Research Findings. *Personality Science*, *2*, 1–20. https://doi.org/10.5964/ps.6043

Letzring, T. D., Wells, S. M., & Funder, D. C. (2006). Information quantity and quality affect the realistic accuracy of personality judgment. *Journal of Personality and Social Psychology*, *91*(1), 111–123. https://doi.org/10.1037/0022-3514.91.1.111

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*(34), 211–222. https://doi.org/10.1024/1010-0652.23.34.211

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

Marsh, H. W. (1992). Content specificity of relations between academic achievement and academic self-concept. *Journal of Educational Psychology*, *84*(1), 35–42. https://doi.org/10.1037/0022-0663.84.1.35

Marsh, H. W., & Craven, R. G. (2006). Reciprocal Effects of Self-Concept and Performance From a Multidimensional Perspective. Beyond Seductive Pleasure and Unidimensional

Perspectives. *Perspectives on Psychological Science*, *1*(2), 133–163. https://doi.org/10.1111/j.1745-6916.2006.00010.x

Möller, J., Pohlmann, B., Köller, O., & Marsh, H. W. (2016). A meta-analytic path analysis of the internal/external frame of reference model of academic achievement and academic self-concept. *Review of Educational Research*, *79*(3), 1129–1167. https://doi.org/10.3102/0034654309337522

Montoya, R. M., & Horton, R. S. (2013). A meta-analytic investigation of the processes underlying the similarity-attraction effect. *Journal of Social and Personal Relationships*, *30*(1), 64–94. https://doi.org/10.1177/0265407512452989

Muntoni, F., & Retelsdorf, J. (2018). Gender-specific teacher expectations in reading—The role of teachers' gender stereotypes. *Contemporary Educational Psychology*, *54*, 212–220. https://doi.org/10.1016/j.cedpsych.2018.06.012

Murphy, N. A., & Hall, J. A. (2021). Capturing Behavior in Small Doses: A Review of Comparative Research in Evaluating Thin Slices for Behavioral Measurement. *Frontiers in Psychology*, *12*. https://doi.org/10.3389/fpsyg.2021.667326

Murphy, N. A., Hall, J. A., & Colvin, C. R. (2003). Accurate intelligence assessments in social interactions: Mediators and gender effects. *Journal of Personality*, *71*(3), 465–493. https://doi.org/10.1111/1467-6494.7103008

Murphy, N. A., Hall, J. A., Mast, M. S., Ruben, M. A., Frauendorfer, D., Blanch-Hartigan, D., Roter, D. L., & Nguyen, L. (2015). Reliability and Validity of Nonverbal Thin Slices in Social Interactions. *Personality and Social Psychology Bulletin*, *41*(2), 199–213. https://doi.org/10.1177/0146167214559902

Murphy, N. A., Hall, J. A., Ruben, M. A., Frauendorfer, D., Schmid Mast, M., Johnson, K. E., & Nguyen, L. (2018). Predictive Validity of Thin-Slice Nonverbal Behavior from Social Interactions. *Personality & Social Psychology Bulletin*, *45*(7), 983–993. https://doi.org/10.1177/0146167218802834

Murphy, M. J., Nelson, D. A., & Cheap, T. L. (1981). Rated and actual performance of high school students as a function of sex and attractiveness. *Psychological Reports*, *48*(1), 103–106. https://doi.org/10.2466/pr0.1981.48.1.103

Nestler, S., & Back, M. D. (2013). Applications and Extensions of the Lens Model to Understand Interpersonal Judgments at Zero Acquaintance. *Current Directions in Psychological Science*, *22*(5), 374–379. https://doi.org/10.1177/0963721413486148

Nestler, S., Egloff, B., Küfner, A. C. P., & Back, M. D. (2012). An integrative lens model approach to bias and accuracy in human inferences: Hindsight effects and knowledge updating in personality judgments. *Journal of Personality and Social Psychology*, *103*(4), 689–717. https://doi.org/10.1037/a0029461

Nickerson, R. S. (1998). Confirmation bias: A ubiquitous phenomenon in many guises. *Review of General Psychology*, *2*(2), 175–220. https://doi.org/10.1037/1089-2680.2.2.175

Oerke, B., McElvany, N., Ohle, A., Ullrich, M., & Horz, H. (2015). Verbessert sich die diagnostische Urteilsgenauigkeit von Lehrkräften bei längerem Kontakt mit der Klasse? [Does diagnostic accuracy of teachers improve with longer contact to the class?] *Psychologie in Erziehung und Unterricht*, *63*(1), 34-47. https://doi.org/10.2378/peu2016.art04d

Funder, D. C., & Ozer, D. J. (2019). Evaluating Effect Size in Psychological Research: Sense and Nonsense: *Advances in Methods and Practices in Psychological Science*, *2*(2), 156–168. https://doi.org/10.1177/2515245919847202

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*(6), 936–943. https://doi.org/10.1177/0021934705285955

Pesu, L., Viljaranta, J., & Aunola, K. (2016). The role of parents' and teachers' beliefs in children's self-concept development. *Journal of Applied Developmental Psychology*, *44*, 63–71. https://doi.org/10.1016/j.appdev.2016.03.001

Pielmeier, M., Huber, S., & Seidel, T. (2018). Is teacher judgment accuracy of students' characteristics beneficial for verbal teacher-student interactions in classroom? *Teaching and Teacher Education*, *76*, 255–266. https://doi.org/10.1016/j.tate.2018.01.002

Pohl, R., Eisenhauer, M., & Hardt, O. (2003). SARA: A cognitive process model to simulate the anchoring effect and hindsight bias. *Memory*, *11*(4–5), 337–356. https://doi.org/10.1080/09658210244000487

Pohl, R. F. (2007). Ways to Assess Hindsight Bias. *Social Cognition*, *25*(1), 14–31. https://doi.org/10.1521/soco.2007.25.1.14

Praetorius, A.-K., Berner, V.-D., Zeinz, H., Scheunpflug, A., & Dresel, M. (2013). Judgment Confidence and Judgment Accuracy of Teachers in Judging Self-Concepts of Students. *The Journal of Educational Research*, *106*(1), 64–76. https://doi.org/10.1080/00220671.2012.667010

Praetorius, A.-K., Drexler, K., Rösch, L., Christophel, E., Heyne, N., Scheunpflug, A., Zeinz, H., & Dresel, M. (2015). Judging students' self-concepts within 30s? Investigating judgement accuracy in a zero-acquaintance situation. *Learning and Individual Differences*, *37*, 231–236. https://doi.org/10.1016/j.lindif.2014.11.015

Praetorius, A.-K., Kastens, C., Hartig, J., & Lipowsky, F. (2016). Haben Schüler mit optimistischen Selbsteinschätzungen die Nase vorn? Zusammenhänge zwischen optimistischen, realistischen und pessimistischen Selbstkonzepten und der Leistungsentwicklung von Grundschulkindern [Are students with optimistic self-concepts one step ahead? Relations between optimistic, realistic, and pessimistic self-concepts and the achievement development of primary school children]. *Zeitschrift für Entwicklungspsychologie und Pädagogische Psychologie*, *48*(1), 14–26. https://doi.org/10.1026/0049-8637/a000140

Pretsch, J., Flunger, B., Heckmann, N., & Schmitt, M. (2013). Done in 60 s? Inferring teachers' subjective well-being from thin slices of nonverbal behavior. *Social Psychology of Education*, *16*(3), 421–434. https://doi.org/10.1007/s11218-013-9223-9

Pretzlik, U., Olsson, J., Nabuco, M. E., & Cruz, I. (2003). Teachers' implicit view of intelligence predict pupils' self-perception as learners. *Cognitive Development*, *18*(4), 579–599. https://doi.org/10.1016/j.cogdev.2003.09.008

Protivínský, T., & Münich, D. (2018). Gender Bias in teachers' grading: What is in the grade. *Studies in Educational Evaluation*, *59*, 141–149. https://doi.org/10.1016/j.stueduc.2018.07.006

Rausch, T., Karing, C., Dörfler, T., & Artelt, C. (2015). Personality similarity between teachers and their students influences teacher judgement of student achievement. *Educational Psychology*, *36*(5), 1–16. https://doi.org/10.1080/01443410.2014.998629

Ready, D. D., & Wright, D. L. (2011). Accuracy and Inaccuracy in Teachers' Perceptions of Young Children's Cognitive Abilities: The Role of Child Background and Classroom

Context. *American Educational Research Journal*, *48*(2), 335–360. https://doi.org/10.3102/0002831210374874

Reiss, K., Sälzer, C., Schiepe-Tiska, A., Klieme, E., & Köller, O. (2016). *PISA 2015 – Eine Studie zwischen Kontinuität und Innovation* [PISA – A study between continuity and innovation]. Waxmann.

Ritts, V., Patterson, M. L., & Tubbs, M. E. (1992). Expectations, Impressions, and Judgments of Physically Attractive Students: A Review. *Review of Educational Research*, *62*(4), 413–426. https://doi.org/10.3102/00346543062004413

Robinson-Cimpian, J. P., Lubienski, S. T., Ganley, C. M., & Copur-Gencturk, Y. (2014). Teachers' perceptions of students' mathematics proficiency may exacerbate early gender gaps in achievement. *Developmental Psychology*, *50*(4), 1262–1281. https://doi.org/10.1037/a0035073

Rothchild, J. (2007). Gender Bias. In *The Blackwell Encyclopedia of Sociology*. John Wiley & Sons, Ltd. https://doi.org/10.1002/9781405165518.wbeosg011

Rubie-Davies, C., Hattie, J., & Hamilton, R. (2006). Expecting the best for students: Teacher expectations and academic outcomes. *British Journal of Educational Psychology*, *76*(3), 429–444. https://doi.org/10.1348/000709905X53589

Rubie-Davies, C. M. (2010). Teacher expectations and perceptions of student attributes: Is there a relationship? *The British Journal of Educational Psychology*, *80*(1), 121–135. https://doi.org/10.1348/000709909X466334

Schön, D. A. (1987). Schon, D. A. (1987). *Educating the Reflective Practitioner. Toward a New Design for Teaching and Learning in the Professions.* Jossey-Bass Publishers.

Schön, D. A. (2017). *The Reflective Practitioner: How Professionals Think in Action*. Routledge. https://doi.org/10.4324/9781315237473

Schrader, F.-W. (1989). *Diagnostische Kompetenz von Lehrern und ihre Bedeutung für die Gestaltung und Effektivität des Unterrichts* [Diagnostic competence of teachers and its' role for lesson planing and lesson efficacy]. Lang.

Schrader, F.-W. (2013). Diagnostische Kompetenz von Lehrpersonen [Diagnostic competence of teachers]. *Beiträge zur Lehrerbildung*, *31*(2)*,* 154–165. https://doi.org/10.25656/01:13843

Schrader, F.-W., & Helmke, A. (1987). Diagnostische Kompetenz von Lehrern. Komponenten und Wirkungen [Diagnostic competence of teachers. Components and effects]. In *Empirische Pädagogik*, *1*(1), 27–52.

Schrader, F.-W., & Helmke, A. (2014). Alltägliche Leistungsbeurteilung durch Lehrer [Teachers' everyday performance evaluations]. In F. E. Weinert (Ed.), *Leistungsmessungen in Schulen* (pp. 45–58). Beltz.

Schrader, F.-W., & Helmke, A. (2015). School Achievement: Motivational Determinants and Processes. In J. D. Wright (Ed.), *International Encyclopedia of the Social & Behavioral Sciences* (2nd ed., pp. 48–54). Elsevier. https://doi.org/10.1016/B978-0-08-097086-8.26055-8

Seidel, T., & Shavelson, R. J. (2007). Teaching effectiveness research in the past decade: The role of theory and research design in disentangling meta-analysis results. *Review of Educational Research*, *77*(4), 454–499. https://doi.org/10.3102/0034654307310317

Shavelson, R. J., Hubner, J. J., & Stanton, G. C. (1976). Self-Concept: Validation of Construct Interpretations. *Review of Educational Research*, *46*(3), 407–441. https://doi.org/10.3102/00346543046003407

Shoda, Y. (1999). A unified framework for the study of behavioral consistency: Bridging person × situation interaction and the consistency paradox. *European Journal of Personality*, *13*(5, Spec Issue), 361–387. https://doi.org/10.1002/(SICI)1099-0984(199909/10)13:5<361::AID-PER362>3.0.CO;2-X

Solga, H., & Pfahl, L. (2009). Doing Gender Im Technisch-Naturwissenschaftlichen Bereich. In J. Milberg (Ed.), *Förderung des Nachwuchses In Technik und Naturwissenschaft: Beiträge zu den Zentralen Handlungsfeldern* (pp. 155–218). Springer. https://doi.org/10.1007/978-3-642-01123-8_4

Spinath, B. (2005). Akkuratheit der Einschätzung von Schülermerkmalen durch Lehrer und das Konstrukt der diagnostischen Kompetenz [Accuracy of Teacher Judgments on Student Characteristics and the Construct of Diagnostic Competence]. *Zeitschrift Für Pädagogische Psychologie*, *19*(1), 85–95. https://doi.org/10.1024/1010-0652.19.12.85

Stang-Rabrig, J., & Urhahne, D. (2016). Wie gut schätzen Lehrkräfte Leistung, Konzentration, Arbeits- und Sozialverhalten ihrer Schülerinnen und Schüler ein? Ein Beitrag zur diagnostischen Kompetenz von Lehrkräften [How do teachers rate students' achievement,

attention, work habits and social behavior? A contribution to the diagnostic competence of teachers]. *Psychologie in Erziehung Und Unterricht*, *63*(3), 204–219. https://doi.org/10.2378/peu2016.art18d

Stewart, M. (1998). Gender issues in physics education. *Educational Research*, *40*(3), 283–293. https://doi.org/10.1080/0013188980400302

Stopfer, J. M., Egloff, B., Nestler, S., & Back, M. D. (2014). Personality Expression and Impression Formation in Online Social Networks: An Integrative Approach to Understanding the Processes of Accuracy, Impression Management and Meta–Accuracy. *European Journal of Personality*, *28*(1), 73–94. https://doi.org/10.1002/per.1935

Südkamp, A., Kaiser, J., & Möller, J. (2012). Accuracy of teachers' judgments of students' academic achievement: A meta-analysis. *Journal of Educational Psychology*, *104*(3), 743–762. https://doi.org/10.1037/a0027627

Südkamp, A., Möller, J., & Pohlmann, B. (2008). Der Simulierte Klassenraum. *Zeitschrift Für Pädagogische Psychologie*, *22*(34), 261–276. https://doi.org/10.1024/1010-0652.22.34.261

Südkamp, A., & Praetorius, A.-K. (2017). Eine Einführung in das Thema der diagnostischen Kompetenz von Lehrkräften [An introduction to the topic of teachers' diagnostic competence]. In A. Südkamp & A. – K. Praetorius (Eds.), *Diagnostische Kompetenz von Lehrkräften: Theoretische und methodische Weiterentwicklungen* (pp. 13–18). Waxmann Verlag.

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

Szumski, G., & Karwowski, M. (2019). Exploring the Pygmalion effect: The role of teacher expectations, academic self-concept, and class context in students' math achievement. *Contemporary Educational Psychology*, *59*, 101787. https://doi.org/10.1016/j.cedpsych.2019.101787

Timmermans, A. C., Kuyper, H., & van der Werf, G. (2015). Accurate, inaccurate, or biased teacher expectations: Do Dutch teachers differ in their expectations at the end of primary education? *British Journal of Educational Psychology*, *85*(4), 459–478. https://doi.org/10.1111/bjep.12087

Timmermans, A. C., & Rubie-Davies, C. M. (2018). Do teachers differ in the level of expectations or in the extent to which they differentiate in expectations? Relations between teacher-level expectations, teacher background and beliefs, and subsequent student performance. *Educational Research and Evaluation*, *24*(3–5), 241–263. https://doi.org/10.1080/13803611.2018.1550837

Timmermans, A. C., Rubie-Davies, C. M., & Wang, S. (2021). Adjusting expectations or maintaining first impressions? The stability of teachers' expectations of students' mathematics achievement. *Learning and Instruction*, *75*, 101483. https://doi.org/10.1016/j.learninstruc.2021.101483

Upadyaya, K., & Eccles, J. (2015). Do teachers' perceptions of children's math and reading related ability and effort predict children's self-concept of ability in math and reading? *Educational Psychology*, *35*(1), 110–127. https://doi.org/10.1080/01443410.2014.915927

Urhahne, D., Chao, S.-H., Florineth, M. L., Luttenberger, S., & Paechter, M. (2011). Academic self-concept, learning motivation, and test anxiety of the underestimated student. *The British Journal of Educational Psychology*, *81*(1), 161–177. https://doi.org/10.1348/000709910X504500

Urhahne, D., & Wijnia, L. (2021). A review on the accuracy of teacher judgments. *Educational Research Review*, *32*, 100374. https://doi.org/10.1016/j.edurev.2020.100374

Veenman, M. V. J., Van Hout-Wolters, B. H. A. M., & Afflerbach, P. (2006). Metacognition and learning: Conceptual and methodological considerations. *Metacognition and Learning*, *1*(1), 3–14. https://doi.org/10.1007/s11409-006-6893-0

Wang, M.-T., & Degol, J. L. (2017). Gender Gap in Science, Technology, Engineering, and Mathematics (STEM): Current Knowledge, Implications for Practice, Policy, and Future Directions. *Educational Psychology Review*, *29*(1), 119–140. https://doi.org/10.1007/s10648-015-9355-x

Wang, S., Rubie-Davies, C. M., & Meissel, K. (2018). A systematic review of the teacher expectation literature over the past 30 years. *Educational Research and Evaluation*, *24*(3–5), 124–179. https://doi.org/10.1080/13803611.2018.1548798

Wang, S., Rubie-Davies, C. M., & Meissel, K. (2020). The stability and trajectories of teacher expectations: Student achievement level as a moderator. *Learning and Individual Differences*, *78*, 101819. https://doi.org/10.1016/j.lindif.2019.101819

Wigfield, A., & Karpathian, M. (1991). Who Am I and What Can I Do? Children's Self-Concepts and Motivation in Achievement Situations. *Educational Psychologist*, *26*(3–4), 233–261. https://doi.org/10.1080/00461520.1991.9653134

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

Zimmerman, B. J. (1995). Self-regulation involves more than metacognition: A social cognitive perspective. *Educational Psychologist*, *30*(4), 217–221. https://doi.org/10.1207/s15326985ep3004_8

Zohar, A., & Bronshtein, B. (2005). Physics teachers' knowledge and beliefs regarding girls' low participation rates in advanced physics classes. *International Journal of Science Education*, *27*(1), 61–77. https://doi.org/10.1080/0950069032000138798

APPENDIX

MANUSCRIPT 1

Urteilsakkuratheit von Lehrkräften bei der Einschätzung nicht-kognitiver Variablen im Fach Physik:

Geschlechtsbezogene Verzerrungseffekte

Caroline Verena Bhowmik[1], Justine Stang-Rabig[2], Katharina Hellmann[3] & Friedrich-Wilhelm Schrader[1]

[1] Rheinland-Pfälzische Technische Universität Kaiserlautern-Landau (RPTU), Institut für Psychologie

[2] TU Dortmund, Institut für Schulentwicklungsforschung

[3] Pädagogische Hochschule Freiburg, Institut für Erziehungswissenschaft

Zusammenfassung

Die Urteilsakkuratheit ist eine wichtige Determinante professioneller Handlungs-kompetenz von Lehrkräften. Lehrkrafturteile können jedoch von Verzerrungseffekten beeinflusst werden, deren Gründe bislang noch wenig beforscht sind. Untersucht wurde, welche Rolle das Geschlecht von Lernenden bei der Einschätzung nicht-kognitiver Variablen im Fach Physik spielt. In zwei Studien wurden Lehramtsstudierende (Studie 1; $N = 27$) und Lehrkräfte (Studie 2; $N = 10$) gebeten, $N = 207$ Lernende hinsichtlich drei zentraler Merkmale von Leistungsmotivation (Fähigkeitsselbstkonzept (FSK), wahrgenommene Aufgabenschwierigkeit (WS), Anstrengungs-bereitschaft (AB)) zu beurteilen. Lehramtsstudierende schätzten das FSK mit mittlerer Genauigkeit ein, die Beurteilung der anderen Merkmale fiel ihnen schwerer. Die Urteilsakkuratheit für das FSK fiel zudem bei den Jungen höher aus. FSK und WS wurde von den Lehramtsstudierenden zusätzlich im Sinne eines Gender Bias signifikant verschätzt. Auch Lehrkräfte schätzten das FSK mit mittlerer Genauigkeit ein. Die Einschätzung war zudem für die Jungen genauer. Implikationen für Forschung sowie Schul- und Unterrichtspraxis werden diskutiert.

Schlüsselwörter: Diagnostische Kompetenz, Fähigkeitsselbstkonzept, Gender Bias, Physik, Urteilsakkuratheit

Summary

Judgment accuracy is an important determinant of teachers' professional competence. However, teacher judgment accuracy can be influenced by various judgmental biases, whose underlying conditions are to date mainly unknown. We investigated the role of students' gender for teachers' assessment of non-cognitive variables in physics. In two studies, teacher students (study 1; $N = 27$) and experienced teachers (study 2; $N = 10$) judged students ($N = 207$) on the three non-cognitive characteristics academic self-concept, perceived task difficulty, and willingness to make an effort. Teacher students achieved on average medium levels of judgment accuracy regarding the academic self-concept, with accuracy being higher for boys. Other characteristics were assessed less accurately. Moreover, we found possible gender-specific judgemental biases of students' academic self-concept and perceived task difficulty that point towards a gender bias. Experienced teachers also achieved medium accuracy regarding students' self-concept and displayed more difficulties in assessing girls' characteristics accurately. Implications for research and education are discussed.

Keywords: academic self-concept, diagnostic competence, gender bias, judgment accuracy, physics

1 Einleitung

Die Akkuratheit, mit der Lehrkräfte Schülerleistungen oder -merkmale einschätzen, ist für erfolgreiche Lehr-Lernprozesse von Lernenden von großer Relevanz. Diagnostische Urteile können dabei sowohl von Merkmalen auf Seiten des Urteilers (z.B. Professionswissen, Terhart, 2009) als auch auf Seiten des Beurteilten (z.B. ethnischer Hintergrund, Maaz et al., 2008) beeinflusst werden. In der Güte der Urteilsakkuratheit bestehen entsprechend oftmals große interindividuelle Unterschiede (z. B. Hoge und Coladarci, 1989; Südkamp, Kaiser und Möller, 2012). Ein relevanter Faktor, der von Seiten der Beurteilten die Urteilsgenauigkeit beeinflussen kann, ist das Geschlecht von Lernenden. In der Schulpraxis treten relevante geschlechtsbezogene Verzerrungseffekte von Urteilen regelmäßig auf (z.B. Südkamp et al., 2012) und kommen insbesondere in den MINT-Fächern (Mathematik, Informatik, Naturwissenschaft, Technik) zum Tragen (z.B. Hofer, 2015; Kuhl und Hannover, 2012). Im Rahmen der vorliegenden Arbeit wird in Anlehnung an die drei Akkuratheitskomponenten nach Schrader (1989) die Rolle eines geschlechtsbezogenen Verzerrungseffektes für die Urteilsgenauigkeit hinsichtlich nicht-kognitiver Schülermerkmale untersucht. Anhand von zwei Studien geht die Arbeit der Frage nach, inwiefern das Geschlecht der zu beurteilenden Lernenden die Urteilsakkuratheit von Lehramtsstudierenden (Studie 1) und erfahrenen Lehrkräften (Studie 2) im Kontext des Fachs Physik beeinflusst. Dabei fokussiert die Arbeit die drei leistungsrelevanten motivational-affektiven Schülermerkmale Fähigkeitsselbstkonzept, wahrgenommene Aufgabenschwierigkeit und Anstrengungsbereitschaft.

2 Theoretischer Rahmen

2.1 Zur Bedeutung der diagnostischen Kompetenz von Lehrkräften

Die diagnostische Kompetenz, verstanden als die Fähigkeit, Schülerleistungen und andere Schülermerkmale angemessen zu beurteilen, wird in der Unterrichts- und Schulforschung als zentrale Komponente professioneller Handlungskompetenz von Lehrkräften und als wichtige Determinante erfolgreichen Lehrens und Lernens angesehen (Baumert und Kunter, 2006; KMK, 2019; Praetorius und Südkamp, 2017; Autorinnen und Autoren, anonymisiert). Diagnostische Kompetenz wird zudem als notwendige Bedingung für die Umsetzung adaptiver Lehr-Lernprozesse angesehen (Corno, 2008; Praetorius und Südkamp, 2017; Autorinnen und Autoren, anonymisiert), da eine hohe Adaptivität von Lehrkräften beispielsweise eine Anpassung von

Instruktionen, Aufgaben oder auch Leistungsrückmeldungen an die Lernvoraussetzungen von Lernenden ermöglicht (Beck, Brühwiler und Müller, 2007; Corno, 2008) und hierdurch lernbezogene Bedarfe von Lernenden aufgedeckt und gezielt gefördert werden können. Aufgrund der Bedeutung und vielseitigen Funktionen von Lehrkrafturteilen ist es entsprechend bedeutsam, dass diese Urteile akkurat ausfallen (Alvidrez und Weinstein, 1999; Bailey und Drummond, 2006; Helmke, 2009; Südkamp et al., 2012). Die Akkuratheit, mit der Schülerleistungen und -merkmale beurteilt werden, stellt dabei den Kern diagnostischer Kompetenz dar (Kaiser, Helm, Retelsdorf, Südkamp und Möller, 2012; Autorinnen und Autoren, anonymisiert; Spinath, 2005).

In der Forschung werden diagnostische Urteile von Lehrkräften meist in Bezug auf kognitive Merkmale, insbesondere die Leistung von Lernenden, hin untersucht (z.B. Anders et al., 2010; Karst et al., 2014; Lorenz und Artelt, 2009; Machts, Kaiser, Schmidt und Möller, 2016). So zeigt sich, dass sich akkurate diagnostische Urteile von Lehrkräften positiv auf Lern- und Leistungsvariablen, so beispielsweise auf die Lesekompetenz von Lernenden (Behrmann und Souvignier, 2013; Karing, Pfost und Artelt, 2011), auf klassenbezogene Lernentwicklungen in der Mathematik (Karst, Schoreit und Lipowsky, 2014), auf die individuelle Mathematikleistung (Anders, Kunter, Brunner, Krauss und Baumert, 2010) oder den Fremdspracherwerb (Zhu, Urhahne und Rubie-Davies, 2018) auswirken. Dabei werden jene positiven Auswirkungen teils über adäquat eingesetzte instruktionale Erklärungen (Behrmann und Souvignier, 2013; Autorinnen und Autoren, anonymisiert) oder über die Nutzung kognitiv aktivierender Aufgaben (Anders et al., 2010) vermittelt.

Neben leistungsbezogenen Merkmalen von Lernenden bilden auch nicht-kognitive lern- und leistungsrelevante Merkmale in verschiedenen Domänen einen wichtigen Gegenstand von Lehrkrafturteilen (KMK, 2019; Lipnevich und Roberts, 2012; Möller, Pohlmann, Köller und Marsh, 2009; Südkamp et al., 2012). Um Lernende subjektiv weder zu über- noch zu unterfordern, ist es hilfreich, wenn Lehrkräfte sich auch an motivationalen Lernvoraussetzungen wie beispielsweise dem Fähigkeitsselbstkonzept (FSK), der wahrgenommene Aufgabenschwierigkeit (WS) und der Anstrengungsbereitschaft (AB) orientieren. Das FSK als eine zentrale nicht-kognitive Variable wird als mentale Repräsentation bezüglich der eigenen schulischen bzw. akademischen Fähigkeiten definiert (z. B. Marsh und Craven, 1997; Schöne, 2018). Es gilt als wichtige Determinante schulischer Leistung (Wigfield und Eccles, 2002) und steht in direktem oder indirektem empirischen Zusammenhang mit Leistungen in verschiedenen Fächern und

Disziplinen (z. B. Autorinnen und Autoren, anonymisiert, 2010; Khalaila, 2014; Marsh, Trautwein, Lüdtke, Köller und Baumert, 2005; Pielmeier, Huber und Seidel, 2018; Seidel, 2006; Seidel et al., 2002). Eng verwandt mit dem FSK sind die AB (Efklides und Tsioara, 2002; Schmitz und Skinner, 1993) und die WS von Lernenden, die gemeinsam einen für Leistungsmotivation und -bereitschaft wichtigen Variablenverbund bilden (Metallidou und Efklides, 2001). WS und AB haben Einfluss darauf, wie Lernende eine Aufgabe bearbeiten (Metallidou und Efklides, 2001). So beispielsweise kann die adäquate Auswahl von motivierendem Lernmaterial mit hohem Lebensweltbezug das Interesse an der Sache steigern und die Lernenden dazu veranlassen, sich stärker anzustrengen. Die Anstrengungsbereitschaft hängt darüber hinaus nicht nur von der tatsächlichen Schwierigkeit einer Aufgabe ab, sondern auch davon, ob die Aufgabe von den Lernenden als zu schwer oder auch zu leicht wahrgenommen wird. Beim FSK in der Grundschule wiederum konnte nachgewiesen werden, dass sich insbesondere eine Unterschätzung der eigenen Leistungsfähigkeit zu Beginn der Schulzeit negativ auf die Leistung auswirkt (Praetorius et al., 2016).
Eine akkurate Wahrnehmung der Ausprägung dieser Schülermerkmale durch Lehrkräfte ist daher wichtig, um bei Bedarf gezielt eingreifen und gegensteuern zu können und damit einen auf die Lernvoraussetzungen der Lernenden abgestimmten und der Lernentwicklung förderlichen adaptiven Unterricht zu realisieren.

2.2 Genauigkeit von Lehrkrafturteilen - Stand der Forschung

Zur Bestimmung der Urteilsakkuratheit können verschiedene Indikatoren herangezogen werden. Schrader (1989) unterscheidet, basierend auf der Arbeit von Cronbach (1955), zwischen der Rang-, Niveau- und Differenzierungskomponente. Das zentrale Maß der Urteilsakkuratheit stellt die Rangkomponente dar (Südkamp, Möller und Pohlmann, 2008). Diese gibt an, inwieweit Lehrkräfte die Rangreihe von Lernenden in Bezug auf das zu beurteilende Merkmal, z. B. Leistung oder Fähigkeitsselbstkonzept (FSK), akkurat vorhersagen können (Praetorius, Lipowsky und Karst, 2012). Die Rangkomponente kann somit als Indikator dafür angesehen werden, inwiefern Lehrkräfte die relativen Unterschiede zwischen ihren Lernenden, z. B. in den Leistungen, akkurat diagnostizieren können. Aufgrund ihrer Bedeutung im Variablenverbund zu Leistungsmotivation und Leistungsbereitschaft wurde das FSK in empirischen Arbeiten zur Urteilsakkuratheit von Lehrkräften am häufigsten untersucht (vgl. Urhahne, Timm, Zhu und Tang, 2013).

Bisherige Untersuchungen der Rangkomponente zeigen, dass Lehrkräfte gut in der Lage sind, Schülerleistungen akkurat einzuschätzen. In einer Metaanalyse von Südkamp et al. (2012), in der die Lehrkrafturteilsakkuratheit aus 75 Studien analysiert wurde, zeigte sich eine mittlere Korrelation von r = .63, welche nach Cohen (1988) als großer Effekt interpretiert werden kann. Jedoch zeigte sich auch, dass sich die Akkuratheit von Lehrkrafteinschätzungen zwischen einzelnen Studien (-.03 ≤ r ≤ .83; Südkamp et al., 2012) und zwischen einzelnen Lehrkräften (e.g., .11 ≤ r ≤ .88; Lorenz und Artelt, 2009) stark unterscheidet. Die Niveaukomponente gibt darüber Auskunft, inwieweit eine urteilende Person dazu tendiert, das entsprechende Merkmal einer zu beurteilenden Person zu über- oder zu unterschätzen (Praetorius et al., 2012). In verschiedenen empirischen Arbeiten zeigt sich, dass Lehrkräfte in der Tendenz dazu neigten, Schülerleistungen zu überschätzen (Bates und Nettelback, 2001; Feinberg und Shapiro, 2009; Urhahne et al., 2010). Die Differenzierungskomponente zeigt an, ob ein Urteiler dazu neigt, die Streuung des einzuschätzenden Merkmals zu über- oder zu unterschätzen. Einige Arbeiten berichteten eine Überschätzung der Streuung bei der Beurteilung von kognitiven und nicht-kognitiven Variablen (z.B. von Schülerleistung, -motivation oder dem FSK; Autorinnen und Autoren, anonymisiert; Spinath, 2005; Urhahne et al., 2010), wohingegen andere Arbeiten, z. B. in Bezug auf die Beurteilung der Intelligenz oder Leistungsängstlichkeit, eine Unterschätzung berichten (Spinath, 2005; Südkamp et al., 2008).

Zu Einschätzungen von nicht-kognitiven Merkmalen liegt bislang weniger empirische Evidenz vor als zu Einschätzungen von kognitiven Variablen wie Schülerleistung. Einzelne Studien konnten zeigen, dass es Lehrkräften insgesamt schwerer fällt, nicht-kognitive als kognitive Merkmale akkurat einzuschätzen. So fielen die Korrelationen zwischen Lehrkrafteinschätzungen und Schülerangaben in Bezug auf das FSK bisher niedrig bis mittelhoch aus (.19 ≤ *r* ≤ .43; Praetorius, Karst, Dickhäuser und Lipowsky, 2011; Spinath, 2005; Urhahne et al., 2010). Zudem konnte für verschiedene Bereiche nachgewiesen werden, dass Lehrkräfte das FSK von Lernenden meistens unterschätzen (Mathematik: *M* = -0.16, *SD* = 0.32; Praetorius, Lipowsky und Gollwitzer, 2010; Praetorius, et al., 2011; Schreiben: *M* = -0.15, *SD* = 0.25; Praetorius et al., 2011; Lesen: M = -0.22, *SD* = 0.25; Praetorius et al., 2011). Für das deutlich seltener untersuchte Konstrukt AB werden ebenfalls niedrige Korrelationen zwischen Urteil und tatsächlicher Merkmalsausprägung berichtet (.15 ≤ *r* ≤ .24; Urhahne et al., 2010; Urhahne et al., 2013). Für das nicht-kognitive Merkmal WS liegen bislang keine empirischen Befunde vor. Eine Besonderheit bei der

Untersuchung nicht-kognitiver Variablen im Kontext von Lehrkrafturteilen besteht darin, dass diese Merkmale aufgrund der Fragebogenmethode weniger reliabel gemessen werden können, als dies bei kognitiven Merkmalen, bei denen standardisierte Testverfahren zum Einsatz kommen können, der Fall ist.

Im Fokus der empirischen Forschung standen bislang meist Urteile von erfahrenen Lehrkräften (Pit-ten Cate, Krolak-Schwerdt und Glock, 2016; Stang-Rabrig & Urhahne, 2016; Südkamp et al., 2012). Die Urteilsgenauigkeit von Lehramtsstudierenden, denen in absehbarer Zeit die Aufgabe der Schülerbeurteilung anvertraut wird, wurde hingegen selten untersucht. Kaiser und Möller (2017) untersuchten die Urteilsgenauigkeit von Studierenden zu mehreren Messzeitpunkten anhand des simulierten Klassenraumes. Für die Lehramtsstudierenden ergaben sich dabei für alle Messzeitpunkte mittlere Rangkorrelationen in Bezug auf die einzuschätzende Schülerleistung. Mit Blick auf motivationale Merkmale ließen sich leicht geringere Korrelationen ausfindig machen. Des Weiteren ergab sich, dass die Streuung der Leistung und der Motivation unterschätzt wurde. Zudem wurde auch die Leistung bzw. Motivation der Lernenden über alle Zeitpunkte hinweg überschätzt. Um die Genese von Urteilsgenauigkeit und die entsprechenden Einflussfaktoren besser verstehen zu können, ist somit ein Einbezug von Lehramtsstudierenden in die Forschung zur Urteilsakkuratheit von Lehrkräften zu empfehlen.

2.3 Effekte geschlechtsspezifischer Verzerrungen auf die Urteilsakkuratheit

Da sich in Arbeiten zur Urteilsakkuratheit von Lehrkräften wiederholt gezeigt hat, dass sowohl intra- als auch interindividuelle Unterschiede in der Urteilsakkuratheit bestehen (z. B. Artelt, Stanat, Schneider, und Schiefele, 2001; Krolak-Schwerdt, Böhmer und Gräsel, 2009; Autorinnen und Autoren, anonymisiert; Autorinnen und Autoren, anonymisiert, Autorinnen und Autoren, anonymisiert, Südkamp et al., 2012; Urhahne et al., 2013), wird zunehmend die Frage nach den möglichen Ursachen für diese Unterschiede in den Fokus von Forschungsarbeiten gerückt.

Forschungsarbeiten im klinischen Kontext haben gezeigt, dass Erwartungen oder stereotype Überzeugungen die Auswahl und Integration von Informationen steuern und folglich auch die Urteilsbildung beeinflussen können (Garb, 1997). Stereotype Erwartungen beschreiben generalisierte Erwartungen gegenüber einer bestimmten Gruppe und können bei Lehrkräften zu Verzerrungen in deren Urteilen führen. Dies erfolgt dadurch, dass zur Urteilsbildung auch

Merkmale herangezogen werden, welche nicht zur Urteilsbildung beitragen und diese beeinträchtigen können (z.B. Ditton und Krüsken 2006; Kaiser et al., 2015; Stang und Urhahne, 2016). Urteilsverzerrungen aufgrund von Stereotypen wurden in der Forschung zur Urteilsakkuratheit von Lehrkräften im Schulkontext bislang wenig untersucht (Südkamp et al., 2012). Südkamp und Kollegen (2012) leiten aus ihrem heuristischen Modell der Urteilsakkuratheit jedoch ab, dass insbesondere das Geschlecht der beurteilten Personen, möglicherweise durch einen indirekten Einfluss, urteilsverzerrend wirkt und die Urteilsakkuratheit entsprechend beeinflussen kann. Solch ein geschlechtsbezogener Verzerrungseffekt wurde in verschiedenen Studien nachgewiesen (Kaiser, Südkamp, und Möller, 2017; Parks und Kennedy, 2007; Pit-ten Cate, Krolak-Schwerdt, Glock und Markova, 2014).

Geschlechtsbezogene Verzerrungseffekte aufgrund von stereotypen Erwartungen können daraus resultieren, dass Unterschieden zwischen Gruppen mehr Bedeutung zugesprochen wird als der Variation innerhalb einer Gruppe (Dovidio, Hewstone, Glick und Esses, 2010; Ellemers, 2018). Somit werden Individuen häufig als Repräsentanten einer gesamten Gruppe wahrgenommen (Ellemers, 2018). Von einem Geschlechtsstereotyp spricht man, wenn das Geschlecht von Individuen als Hauptmerkmal einer Gruppe identifiziert wird (Eichler, Fuchs und Maschewski-Schneider, 2000). Stereotype Erwartungen könnten sich als Verzerrungseffekt auch auf die Urteilsakkuratheit von Lehrkräften auswirken und sich daher auch in den Akkuratheitskomponenten zeigen. Die Tendenz zur Über- oder Unterschätzung nicht-kognitiver Merkmale von Lernenden sollte sich im Urteilsniveau (Niveaukomponente) widerspiegeln. Zusätzlich kann sich die Stereotypisierung in einer Homogenisierung der Urteile für einzelne Gruppen und in einem geringeren Unterscheidungsvermögen innerhalb einer Gruppe (Differenzierungs-, Rangkomponente) niederschlagen.

Ein geschlechtsbezogener Verzerrungseffekt wird in der schulischen und unterrichtlichen Praxis vor allem in den sogenannten MINT-Fächern (Mathematik, Informatik, Naturwissenschaft, Technik) angenommen (Leslie, Cimpian, Meyer und Freeland, 2015). Lehrkräfte scheinen häufig davon auszugehen, dass Mädchen und Jungen sich in ihren naturwissenschaftlichen Leistungen unterscheiden, obwohl tatsächlich keine Unterschiede bestehen. Auch die von Lehrkräften eingeschätzte WS von Aufgaben und die AB von Lernenden wird von geschlechtsspezifischen Überzeugungen mitbeeinflusst (Tiedemann, 2000).

Innerhalb der MINT-Fächer kommt, stereotype Erwartungen betreffend, dem Unterrichtsfach Physik eine besondere Bedeutung zu. Physik gilt als genuin maskulines Unterrichtsfach. Dies kann zur Folge haben, dass viele Lehrkräfte vor allem den männlichen Schüler im Blick haben, wenn sie das Fach unterrichten. In einer solchen, androzentrisch geprägten Sichtweise stellen männliche Schüler die Norm dar, so dass der Blick vergleichsweise wenig durch fachinkompatible Stereotype verstellt wird. Daraus kann gefolgert werden, dass Lehrkräfte männliche Schüler tendenziell eher unvoreingenommen und unverzerrter wahrnehmen. Eine androzentrische Sichtweise auf das Fach bedeutet gleichzeitig, dass das Fach als für Mädchen weniger passend wahrgenommen wird und Mädchen als eher ungeeignet für das Fach angesehen werden. Es könnten daher Geschlechtsstereotype mobilisiert werden, die eine genaue Wahrnehmung erschweren. Das könnte wiederum zur Folge haben, dass Schülerinnen bestimmte Eigenschaften zugeschrieben werden, die ihre Eignung für dieses Fach in Frage stellen und einem Erfolg in diesem Fach eher abträglich sind. Dieses Stereotyp verstellt zugleich den Blick auf die Individualität der Lernenden und geht eher von einem Durchschnittstypus aus, so dass es zu einer Homogenisierung der Beurteilung kommen könnte.

Schüler beiderlei Geschlechts sind nicht nur gegenüber den Naturwissenschaften im Allgemeinen (z. B. Osborne, Simon und Collins, 2003), sondern insbesondere dem Unterrichtsfach Physik gegenüber stark negativ eingestellt (Kessels, Rau und Hannover, 2006). Auch das Interesse an Physik sowie die Leistungen in diesem Fach sind oftmals geringer oder über die Lernzeit hinweg abnehmend (z. B. Baumert und Lehmann, 1997; Muckenfuß, 1995; Prenzel, Reiss und Hasselhorn, 2009). Physik als Schulfach ist bei Lernenden deutlich unbeliebter als nicht-naturwissenschaftliche Fächer (Merzyn, 2008, 2009; Muckenfuß, 1995). Zudem wurde mehrfach nachgewiesen, dass die Physik, wie MINT-Fächer im Allgemeinen, als eher maskuline, gleichzeitig aber auch anspruchsvolle Fächer gelten, in welchem Jungen mehr Kompetenzen zugesprochen werden als gleichaltrigen Mädchen (Kessels et al., 2006; Kessels, 2008; Kessels, 2015; Solga und Pfahl, 2009). Der Erfolg von Mädchen in der Physik wird dabei meist auf Anstrengung zurückgeführt, während erfolgreiche Leistungen bei Jungen auf ihre Fähigkeiten und ihr Talent in diesem Bereich attribuiert werden (Kessels, 2015; Lightbody, Siann, Stocks und Walsh, 1996). Wird Mädchen von Lehrkräften eine Begabung in Physik zugesprochen, wird dies von den Mädchen nur bedingt positiv aufgenommen und äußert sich zudem nicht in verstärktem Interesse am Unterrichtsfach (Kessels, Warner, Holle und Hannover, 2008). Diese Tatsache spiegelt sich anscheinend auch im

FSK von Schülerinnen wider, welches im Vergleich zu dem von Schülern stärker negativ ausgeprägt ist (Kessels und Hannover, 2004; Jansen, Schroeders, Lüdtke und Marsh, 2019).
Was die tatsächlichen Leistungen von Schülerinnen in der Physik betrifft, so zeigt sich in der Literatur ein eher uneinheitliches Bild. So konnte in einigen Untersuchungen aufgezeigt werden, dass die tatsächlichen Leistungen von Schülerinnen zum Teil geringer ausfallen als bei männlichen Schülern (Reiss, Sälzer, Schiepe-Tiska, Klieme und Köller, 2016). Dies war auch bei Schülerinnen der Fall, die ansonsten über überdurchschnittliche kognitive Fähigkeiten verfügten (Hofer und Stern, 2016). Diese Befunde stehen jedoch in Kontrast zu anderen Arbeiten, welche zeigen, dass Schülerinnen in naturwissenschaftlichen Fächern besser abschnitten als Schüler (Herwartz-Emden et al., 2012).

Studien, die sich mit geschlechtsbezogenen Verzerrungseffekten in der Einschätzung der Schülerleistung und anderer Schülermerkmale durch Lehrkräfte befassen, wurden bislang meist im Fach Mathematik durchgeführt (z. B. Krauss et al., 2008; Südkamp et al., 2012). So zeigte sich in verschiedenen Arbeiten, dass Mathematiklehrkräfte Jungen bessere Kenntnisse zuschreiben als Mädchen (Kuhl und Hannover, 2012; Jussim und Eccles, 1992; Robinson-Cimpian, Lubienski, Ganley und Copur-Gencturk, 2014; Tiedemann, 2000). Auch für Lehramtsstudierende konnte dies belegt werden (Holder und Kessels, 2007). Auch eine Studie im Fach Physik berichtet einen analogen Effekt: Im Rahmen einer Onlinestudie, in welcher das identische Testergebnis eines fiktiven Schülers bzw. einer fiktiven Schülerin beurteilt werden sollte, zeigte sich ein Gender Bias zuungunsten der Schülerinnen, die schlechter bewertet wurden (Hofer, 2015). Es konnte zudem gezeigt werden, dass Schülerinnen bei gleicher Leistung in Physik im Mittel schlechtere Noten erhielten als männliche Schüler (Hofer, 2015). Bislang fehlen jedoch ausreichend Forschungsergebnisse insbesondere aus der Physik, die einen möglichen Verzerrungseffekt von Geschlechtsstereotypen auf die Urteilsgenauigkeit von Lehrkräften untersuchen.

3 Zielsetzung, Fragestellungen und Hypothesen für Studien eins und zwei

Ziel der vorliegenden Arbeit ist es, geschlechtsbezogene Verzerrungseffekte und ihre Rolle für die Genauigkeit der Urteilsakkuratheit bei nicht-kognitiven Schülermerkmalen in der Physik zu untersuchen. Untersucht wurden Merkmale, die für Leistungsmotivation und Leistungsbereitschaft von Lernenden wichtig sind (FSK, WS, AB). Die Urteilsakkuratheit für diese Merkmale wurde bislang über alle schulischen Domänen hinweg vergleichsweise selten

untersucht. Des Weiteren spielen die ausgewählten Merkmale insbesondere für die Forschung zu geschlechtsbezogenen Verzerrungseffekten in den MINT-Fächern eine zentrale Rolle, da Lehrkräften häufig und implizit eine Ungleichverteilung der Ausprägungen von FSK, WS und AB zwischen Schülerinnen und Schülernzeigen. Dies kann sich in der Erwartung von Lehrkräften niederschlagen, dass sich Schülerinnen in der Physik mehr anstrengen (müssen) und ihnen die Aufgaben in der Physik schwerer fallen als ihren männlichen Mitschülern (vgl. Kessels, 2015). Zudem hat sich gezeigt, dass analog dazu das FSK von Schülerinnen meist geringer ausgeprägt ist als von Schülern. Auf Basis dieser theoretischen Annahmen soll insbesondere der Frage nachgegangen werden, ob das Geschlecht der beurteilten Lernenden einen differenziellen Einfluss auf die Urteilsakkuratheit von angehenden und berufserfahrenen Lehrkräften hat, also ein sogenannter Gender Bias vorliegt. Daraus ergeben sich folgende Fragestellungen und Hypothesen:

1. Wie werden die Schülermerkmale FSK, WS und AB im Hinblick auf grundlegende Komponenten der Urteilsakkuratheit (Rang-, Niveau-, und Differenzierungskomponente) eingeschätzt?

Es wird erwartet, dass die Rangfolge des FSK, der AB und der WS mit mittlerer Genauigkeit eingeschätzt werden kann. Des Weiteren wird angenommen, dass das Niveau des FSK unterschätzt bzw. die Streuung des Merkmals überschätzt wird. Für die WS und AB lassen sich aufgrund des unzureichenden Forschungsstands keine weiteren Vorhersagen treffen.

2. Werden das allgemeine Merkmalsniveau bzw. Merkmalsunterschiede bei Jungen und Mädchen unterschiedlich hoch und unterschiedlich genau eingeschätzt?

Die bislang gefundenen Urteilsverzerrungen aufgrund möglicher stereotyper Erwartungen von Lehrkräften legen nahe, dass Mädchen homogener wahrgenommen werden als Jungen. Dies hat zur Folge, dass die Genauigkeit der Wahrnehmung von Unterschieden (Rangordnungskomponente) und die relative Urteilsstreuung (Differenzierungskomponente) bei Mädchen geringer ausfallen sollte als bei Jungen. Dementsprechend wird davon ausgegangen, dass sowohl Rangordnungs- als auch Differenzierungskomponente für Mädchen geringer ausfallen als für Jungen. Außerdem wird angenommen, dass Mädchen im Vergleich zu Jungen im Hinblick auf leistungsförderliche motivationale Faktoren eher unterschätzt werden. Für die Komponenten der

Urteilsakkuratheit bedeutet dies, dass bei Mädchen die Niveaukomponente des FSK und der AB niedriger, und die der WS höher ausgeprägt sein sollte als bei Jungen.

4 Studie 1

4.1 Stichprobe und Design

An der Studie nahmen N = 27 Lehramtsstudierende (46.2 % weiblich, M = 24.8 Jahre, SD = 4.06) als Urteilende teil. Die Lehramtsstudierenden waren zum Zeitpunkt der Datenerhebung mit dem Unterrichtsfach Physik im Bachelor of Education eingeschrieben (M = 6. Semester, SD = 2.86) und ihre Teilnahme an der Studie erfolgte im Rahmen eines Physikseminars. Als Referenzwerte für die Urteile wurden die Selbstberichtsdaten von N = 207 Schülern (40.3 % weiblich, M = 15.6 Jahre, SD = 0.63) herangezogen. Die Lernenden von zwei der Schulklassen besuchten zum Zeitpunkt der Studiendurchführung die Realschule, die der restlichen acht Schulklassen das Gymnasium. Die einzelnen Schulklassen besuchten zusammen mit ihren Physiklehrkräften das Physiklabor der Universität. Die Klassenstärke betrug im Mittel 21 Schüler (min. = 17; max. = 23). Die Klassen wurden zunächst in Abhängigkeit von der jeweiligen Klassenstärke auf zwei oder drei Subklassen (insgesamt 27 Subklassen) aufgeteilt: die mittlere Subklassenstärke belief sich auf acht Schüler (SD = 2.04). Für die Bildung der Subklassen wurden die Schulklassen hinsichtlich Geschlecht und Physiknote parallelisiert um sicherzustellen, dass in jeder Subklasse ausreichend Heterogenität hinsichtlich der für die vorliegende Arbeit zentralen Kriterien sichergestellt ist. In 90-minütigen Unterrichtseinheiten arbeiteten die Schüler daraufhin selbstständig in Dyaden an in einem Experimentierheft präsentierten Physikexperimenten zum Thema Solarenergie und wurden dabei von jeweils einem oder einer der 27 Lehramtsstudierenden inhaltlich begleitet. Die Lehramtsstudierenden waren in dieser Zeit dazu angehalten, die Lernenden durch Beobachtung, individualisierte Hilfestellungen und gezielte Nachfragen inhaltlich zu unterstützen. Im Anschluss an die Unterrichtseinheit wurden die Lehramtsstudierenden aufgefordert, ihre Einschätzungen der einzelnen Schüler in ihrer Unterrichtsgruppe hinsichtlich der drei untersuchten Konstrukte abzugeben.

4.2 Material

4.2.1 Schülerselbstberichtsdaten

Das FSK der Schüler im Fach Physik, welches die Gesamtheit der mentalen Repräsentationen der eigenen Fähigkeiten umfasst, wurde anhand einer vier Item umfassenden Skala (α = .88; Bsp.: „Ich bin für Physik begabt"; Seidel, Prenzel, Duit, und Lehrke, 2003) mit vierstufigem Antwortformat (1 = trifft nicht zu, 2 = trifft eher nicht zu, 3 = trifft eher zu, 4 = trifft völlig zu) erfasst. Um valide Indikatoren für die AB, welche darüber Auskunft gibt, wie gut Schüler schulische Anforderungen durch eigene Anstrengung bewältigen können, und für die WS, welche angibt, wie leicht oder schwierig die Lernenden die zu bearbeitenden Aufgaben empfinden, zu erhalten, wurden beide Konstrukte jeweils vor und nach Bearbeitung eines jeden Aufgabenbereiches im Experimentierheft erhoben. Da das Experimentierheft sieben Aufgabenbereiche umfasste, resultierte dies in insgesamt 14 Schülerangaben pro Konstrukt. Die beiden Konstrukte wurden jeweils in einer Pre-Messung („Wie sehr wirst du dich beim Durchführen der Experimente anstrengen?" (AB) und „Was glaubst du, wie leicht / schwer werden dir die Experimente fallen?" (WS); Efklides und Tsiora, 2002) und im Anschluss an den jeweiligen Aufgabenbereich als Post-Messung („Wie sehr hast du dich beim Durchführen der Experimente angestrengt?" (AB) und „Wie leicht / schwer fandest du die Aufgabe?" (WS); Efklides und Tsiora, 2002) anhand eines Items gemessen. Das Antwortformat belief sich auf 1 = sehr leicht, 2 = leicht, 3 = mittel, 4 = schwer und 5 = sehr schwer (WS; α = .94) und 1 = sehr wenig, 2 = wenig, 3 = mittel, 4 = viel und 5 = sehr viel (AB; α = .81). Die Reliabilitäten der eingesetzten Skalen können mit sehr gut beurteilt werden. Für die darauffolgenden Genauigkeitsberechnungen wurden alle Schülerangaben gemittelt und anschließend als Referenzwerte für die Urteile herangezogen.

4.2.2 Urteile der Lehramtsstudierenden und Lehrkräfte

In beiden Studien wurden die Urteile hinsichtlich des FSK der Schüler im Fach Physik anhand eines Items erfasst („Was glauben Sie, wie der Schüler/die Schülerin seine/ihre Fähigkeiten in Physik eingeschätzt hat?"). Das Antwortformat belief sich auf 1 = schlecht bis 5 = gut. Zur Orientierung und Hilfestellung wurde den Urteilenden die von den Lernenden bearbeitete Selbstberichtsskala vorgelegt. Um darüber hinaus der unterschiedlichen Skalierung der Schülerselbstberichtsdaten und der Skala für die Einschätzungen Rechnung zu tragen, wurden die

Urteile vor den weiteren Berechnungen entsprechend des Antwortformats der eingesetzten Schülerskala transformiert. Dazu wurden die Extremwerte der Einschätzungen an die der Schülerselbstberichtsdaten angepasst (Praetorius et al., 2011). Das daraus resultierende, mögliche Minimum der Einschätzungen lag somit bei 1 und das mögliche Maximum bei 4. Die verbleibenden Werte wurden äquidistant zwischen die angepassten Extrempole gesetzt (der ursprüngliche Wert 2 wurde zu 1.75 transformiert, der Wert 3 zu 2.5, der Wert 4 zu 3.25). AB und WS wurden ebenfalls jeweils anhand eines Items eingeschätzt („Wie schätzen Sie Ihren Schüler/Ihre Schülerin ein? Wie sehr strengt er/sie sich normalerweise beim Durchführen von Experimenten an?“ und „Wie schätzen Sie Ihren Schüler/Ihre Schülerin ein? Wie leicht/schwer fallen ihm/ihr normalerweise Experimente im Physikunterricht?“). Das Antwortformat für alle Einschätzungen belief sich analog zu den Schülerselbstberichtsskalen auf 1 = sehr leicht bis 5 = sehr schwer (WS) und 1 = sehr wenig bis 5 = sehr viel (AB).

4.3 Statistische Analysen

Im Rahmen der ersten Fragestellung wurden zur Ermittlung der Rangkomponente Pearson-Korrelationen zwischen Einschätzungen und Schülerangabe klassen- bzw. gruppenweise berechnet und Fisher-Z-transformiert. Die einzelnen Werte wurden gemittelt und rücktransformiert, um den Mittelwert der Rangkomponente als Korrelationskoeffizienten darstellen zu können. Mittels *t*-Test wurde basierend auf den Fisher-*Z*-transformierten Werten zusätzlich ermittelt, ob die mittleren Rangkomponenten für die drei untersuchten Schülermerkmale signifikant von 0 verschieden sind. Die Niveaukomponente wurde als klassen- bzw. gruppenweise berechnete Differenz aus Angaben der Lehrkräfte bzw. Studierende und denen der Schüler gebildet. Werte kleiner 0 bedeuten eine Unterschätzung, wohingegen Werte größer 0 auf eine Überschätzung des Merkmals hindeuten. Die Differenzierungskomponente berechnet sich aus der klassen- bzw. gruppenweise berechneten Streuung der Urteile der Beurteilenden geteilt durch die der Schülerangaben des jeweiligen Merkmals. Werte kleiner 1 deuten auf eine Unterschätzung hin, wohingegen Werte größer 1 auf eine Überschätzung hindeuten. Mittels eines t-Tests für eine Stichprobe wurden die Werte der Niveaukomponente gegen 0 und die der Differenzierungskomponente gegen 1 getestet.

Zur Beantwortung der zweiten Fragestellung wurden zunächst, wie bei der ersten Fragestellung, jeweils die Rang-, Niveau- und Differenzierungskomponente berechnet, diesmal jedoch getrennt für Mädchen und Jungen. Mittels eines *t*-Tests für den Einstichprobenfall wurden

die Werte der Rang- und Niveaukomponente auch hier wieder gegen 0 und die der Differenzierungskomponente gegen 1 getestet, diesmal jedoch getrennt nach Geschlecht. Zur Prüfung, inwieweit Unterschiede bei den Einschätzungen der Merkmale der Mädchen und Jungen bestehen, wurde der Wilcoxon-Test herangezogen. Es handelt sich dabei um gepaarte Stichproben, da für jeden Urteiler zwei Kennwerte (Mädchen, Jungen) vorliegen.

4.4 Ergebnisse Studie 1

4.4.1 Voranalysen und deskriptive Statistiken

In Tab. 1 sind Mittelwerte und Geschlechtsunterschiede der Schülerselbsteinschätzungen sowie im Vergleich dazu die in beiden Studien erhobenen Beurteilungen der Lehrkräfte und Lehramtsstudierenden dargestellt. *T*-Tests für unabhängige Stichproben ergaben, dass die Physiknoten von Mädchen und Jungen sich nicht signifikant voneinander unterschieden. Gleichwohl zeigten sich jedoch signifikante Unterschiede im physikbezogenen FSK. Mädchen wiesen ein signifikant niedrigeres FSK auf als Jungen. In Bezug auf die AB zeigten sich ebenfalls Unterschiede. Mädchen schätzten ihre AB signifikant höher ein als Jungen. Mädchen beurteilten zudem im Vergleich zu Jungen die zu bearbeitenden Aufgaben als signifikant schwieriger. Die in Tab. 2 dargestellten Korrelationen zeigen, dass die Fremdeinschätzungen der Schülermerkmale durch die LA-Studierenden deutlich miteinander zusammenhängen, während die Zusammenhänge zwischen den Selbsteinschätzungen der Schülerinnen und Schüler eher schwach ausgeprägt sind.

Tabelle 1

Lehrerurteile und Schülerselbsteinschätzungen: Mittelwerte (in Klammern: Standardabweichungen) und Geschlechtsunterschiede

Variable	LA-Stud.	Lehrkräfte	SuS gesamt	Mädchen	Jungen	*t* (*df*) Jungen vs. Mädchen
Physiknote	-	-	2.81 (0.94)	2.82 (0.87)	2.80 (0.99)	-.18 (201)
Fähigkeitsselbstkonzept	2.51 (0.84)	2.79 (0.88)	2.58 (0.73)	2.37 (0.65)	2.74 (0.74)	3.78 (204)***
Anstrengungsbereitschaft	3.56 (0.89)	3.,23 (1.05)	3.45 (0.73)	3.64 (0.60)	3.32 (0.79)	-3.27 (202)**
Wahrgenommene Schwierigkeit	2.98 (0.91)	2.80 (0.93)	2.92 (0.53)	3.06 (0.50)	2.81 (0.53)	-3.51 (203)**

Anmerkungen: Noten sind mit 1 = sehr gut bis 6 = ungenügend kodiert; ** $p < .01$, *** $p < .001$.

Tabelle 2

Interkorrelationen der Selbstberichte der SuS und der Fremdurteile für die relevanten Schülermerkmale

Variable	(1)	(2)	(3)	(4)	(5)	(6)	(7)
Selbstberichte SuS	SuS – SuS				SuS – LA-Stud		
(1) Note Physik	-	-.68**	-.18*	.09	-.20**	-.15*	.17*
(2) FSK		-	.13	-.15*	.35**	.12	-.28**
(3) AB			-	.22**	.03	.10	-.01
(4) WS				-	-.21**	-.08	.18*
Fremdurteile Lehrpersonen	SuS – Lehrkraft				Lehrkraft\LA-Stud.		
(5) FSK	-.62**	.51**	.11	-.10	-	.36**	-.70**
(6) AB	-.48	.31**	.15*	-.11	.48**	-	-.40**
(7) WS	.64**	-.56**	-.12	.09	-.73**	-.55**	-

Anmerkungen: SuS = Schülerinnen und Schüler, FSK = Fähigkeitsselbstkonzept, AB = Anstrengungsbereitschaft, WS = wahrgenommene Schwierigkeit; Werte oberhalb der Diagonale beziehen sich auf Lehramtsstudierende, Werte unterhalb auf Lehrkräfte; Noten sind mit 1 = sehr gut bis 6 = ungenügend kodiert; * $p < .05$, ** $p < .01$.

4.4.2 Genauigkeit der Lehramtsstudierendeneinschätzungen

In Tab. 3 sind die Ergebnisse zur Urteilsgenauigkeit der Lehramtsstudierenden dargestellt. Im Hinblick auf die erste Fragestellung zeigte sich für die Rangkomponente, dass die Einschätzungen der Lehramtsstudierenden mäßig mit den Angaben der Lernenden zusammenhingen. Zur Berechnung der statistischen Kennwerte wurden die Rangkomponenten der einzelnen Urteiler Fischer-Z-transformiert und anschließend wieder zurücktransformiert. Der beträchtliche Range ($r = -.70 \leq r \leq .98$) verdeutlicht, dass zwischen den Lehramtsstudierenden bei der Einschätzung aller drei Bereiche große interindividuelle Unterschiede in der Urteilsakkuratheit bestanden. Mittels t-Tests (Einstichprobenfall) wurde geprüft, ob sich die Mittelwerte der Fisher-Z-transformierten Rangordnungskomponenten von 0 unterscheiden. Die Tests ergaben, dass die Einschätzungen der Lehramtsstudierenden hinsichtlich des FSK ($r = .44$, $t(26) = 5.97$, $p < .01$) und der WS ($r = .27$, $t(26) = 2.55$, $p < .05$) signifikant von 0 abwichen. Die Einschätzungen der AB hingegen waren dagegen niedrig und nicht signifikant von 0 verschieden. Die Berechnung der Niveaukomponente ergab, dass die Lehramtsstudierenden das FSK der Schülerinnen und Schüler signifikant überschätzten ($M = 0.46$, $t(26) = 6.56$, $p < .001$), während die Angaben für AB und WS nicht signifikant ausfielen. Die Tests für die Differenzierungskomponente ergaben, dass die

Lehramtsstu-dierenden bei allen drei Merkmalen die Streuung signifikant überschätzten. Rein deskriptiv betrachtet, also ohne statistische Absicherung der Unterschiede, wurde die Streuung der WS dabei am deutlichsten überschätzt (M = 2.25, $t(26)$ = 3.76), gefolgt von AB (M= 1.78, $t(26)$ = 8.03) und schließlich FSK (M = 1.32, $t(26)$ = 10.55) (vgl. Tab. 3). Insgesamt kann festgehalten werden, dass die Rangkomponente für FSK und WS mäßig genau, die Niveaukompetente im Falle des FSK überschätzt und die Streuungen aller drei Schülermerkmale deutlich überschätzt werden. Unterschätzungen gab es weder bei der Niveau- noch bei der Differenzierungskomponente. Die Daten unterstützen die Hypothese 1 daher nur in Teilen.

4.4.3 Unterschiede in der Genauigkeit der Einschätzung von Merkmalsniveau und Merkmalsunterschieden bei Jungen und Mädchen

In Bezug auf die zweite Fragestellung wurde im ersten Schritt zunächst für Mädchen und Jungen separat geprüft, ob sich die Mittelwerte der Diagnosekennwerte in beiden Gruppen signifikant von Null bzw. im Fall der Differenzierungskomponente von Eins unterscheiden. Im zweiten Schritt wurde dann geprüft, ob sich für die drei Diagnosekennwerte signifikante Unterschiede zwischen Mädchen und Jungen nachweisen lassen. Für die Rangkomponente zeigte sich, dass die mittleren Korrelationen für das FSK bei den Jungen (r = .73, $t(19)$ = 6.7, p < .001) signifikant von 0 abwichen (vgl. Tab. 3).

Signifikante Unterschiede in der Genauigkeit der Einschätzungen von Jungen und Mädchen ließen sich lediglich für das FSK nachweisen (r (Mädchen) = .28 vs. r (Jungen) = 73, z = -3.1, n = 17, p < .01). Die Effektstärke (berechnet anhand z geteilt durch die Wurzel aus n) liegt bei r = .79 und das Ergebnis entspricht nach Cohen (1988) einem starken Effekt. Jungen wurden demnach von den Lehramtsstudierenden genauer beurteilt als Mädchen. Für das FSK konnte somit die zweite Hypothese bestätigt werden, die auf der Annahme beruht, dass die Rangkomponente für Mädchen geringer ausfallen sollte als für die Jungen.

Bei der getrennten Berechnung der Niveaukomponente für Mädchen und Jungen ergab sich, dass die Lehramtsstudierenden die WS (M = 0.38, $t(22)$ = 3.28, p < .01) bei den Mädchen signifikant überschätzten. Bei den Jungen wurden dagegen das FSK (M = 0.93, $t(24)$ = 6.81, p < .001) und die AB (M = 0.30, $t(24)$ = 2.59, p < 0.5) überschätzt. Beim Vergleich der beiden Geschlechtsgruppen zeigten sich signifikante Unterschiede für die Merkmale WS (M (Mädchen) = 0.38 vs. M (Jungen) = -0.18, z = -2.9, p < 0.01) und FSK (M (Mädchen) = -0.05 vs. (M (Jungen) =

0.93, z = -4.0, p < .001). Hinsichtlich des FSK und der WS konnten diese Ergebnisse somit die zweite Hypothese in Teilen bestätigen.

Tabelle 3

Rang-, Niveau- und Differenzierungskomponente Studie 1 (Lehramtsstudierende) für die Gesamtstichprobe sowie Substichproben der Mädchen und Jungen

	Gesamt		Mädchen		Jungen		Mädchen vs. Jungen
	M	*SD*	*M*	*SD*	*M*	*SD*	M_{diff}
Rangkomponente							
FSK	.44 **	0.33	.28	0.69	.73***	0.27	-.45**
AB	.03	0.44	-.10	0.75	-.14	0.67	-.04*
WS	.27*	0.34	.30*	0.71	.07	0.60	.23
Niveaukomponente							
FSK	0.46***	0.36	-0.05	0.57	0.93***	0.71	-0.98***
AB	0.07	0.48	0.05	0.53	0.30*	0.60	0.25
WS	0.03	0.36	0.38**	0.59	-0.18	0.50	0.56**
Differenzierungs-komponente							
FSK	1.32 **	0.66	1.21	1.27	0.91	0.60	0.30
AB	1.78 **	1.15	1.62**	1.39	1.98*	2.20	-0.36
WS	2.25 **	1.19	2.50**	2.48	1.88**	1.21	0.62

Anmerkungen. Testungen gegen 0 bzw. 1 mittels *t*-Test;
Vergleich Mädchen vs. Jungen: Wilcoxon-Test * p < .05, ** p < .01, *** p < .001

Die Streuung wurde im Mittel von den Lehramtsstudierenden bei den Mädchen sowohl hinsichtlich der AB (M = 1.62, t(24) = 2.3, p < .01), als auch der WS (M = 2.50, t(22) = 2.9, p < .01) signifikant überschätzt. Somit wich der Quotient der gruppenweisen Streuung der Urteile der Lehramtsstudierenden und der tatsächlichen Streuung signifikant vom Idealwert 1 ab. Auch bei den Jungen wurde sowohl die Streuung der AB (M = 1.98, t(24) = 2.2), p < .05), als auch der WS

($M = 1.88$, $t(23) = 3.6$, $p < .01$) signifikant überschätzt. Mit Blick auf die Fragestellung ergab sich, dass sich die Einschätzung der Streuung der Merkmale nicht signifikant für Mädchen und Jungen unterschieden. Die zweite Hypothese konnte somit anhand der Ergebnisse zur Differenzierungskomponente nicht bestätigt werden.

4.5 Ergebnisdiskussion Studie 1

Im Fokus der vorliegenden Arbeit stand die Urteilsakkuratheit von Lehramtsstudieren-den im Fach Physik hinsichtlich weniger stark beforschter Schülercharakteristika wie dem physikbezogenen FSK, der WS und der AB. Dabei wurde die Urteilsakkuratheit sowohl allgemein, als auch nach Geschlecht der beurteilten Lernenden, untersucht. Als Ursache für inakkurate Urteile im Sinne eines Gender Bias wurde geschlechtsspezifische Urteilsver-zerrungen angenommen.
Zwischen Mädchen und Jungen bestanden in der Physiknote keine Unterschiede, wohl aber in der Selbsteinschätzung des FSK (Kessels und Hannover, 2004) sowie der WS und der AB. Mädchen attestierten sich ein geringeres FSK als Jungen, empfanden im Vergleich zu den Jungen die zu bearbeitenden Aufgaben als schwieriger und schrieben sich eine höhere Anstrengungsbereitschaft zu als Jungen.

Für die Einschätzungen der Lehramtsstudierenden bezüglich FSK und WS wurden mäßig hohe Korrelationen ermittelt, welche in Einklang mit den bisher in der Forschung ermittelten Ergebnissen für nicht-kognitive Merkmale stehen (Praetorius et al., 2011; Spinath, 2005; Urhahne et al., 2010). Zudem zeigte sich, dass bei den Lehramtsstudierenden starke interindividuelle Unterschiede in der Urteilsakkuratheit bestanden (Autorinnen und Autoren, anonymisiert; Autorinnen und Autoren, anonymisiert; Urhahne et al., 2013). Die allenfalls mittelhohen Korrelationen könnten dadurch zu erklären sein, dass den urteilenden Personen die entsprechenden Indikatoren der jeweiligen Merkmale nicht hinreichend klar sind (Autorinnen und Autoren, anonymisiert; Autorinnen und Autoren, anonymisiert). Funder (1995, 2012) weist in seinem Realistic Accuracy Model (RAM) auf die Bedeutsamkeit der Beobachtbarkeit des zu beurteilenden Merkmals hin. Dies bedeutet, dass für akkurate Urteile vom Beurteilten relevante Hinweisreize ausgesendet werden und dem Urteilenden zur Verfügung stehen müssen. Des Weiteren muss der Urteilende in der Lage sein, diese Reize zu erkennen und entsprechend für die Urteilsgenerierung zu nutzen. Zudem ist es bei nicht-kognitiven Variablen für den Urteilenden oftmals nicht möglich, die eigene Einschätzung mit den tatsächlichen Merkmalsausprägungen der Schüler abzugleichen,

da diese via Schülerselbstbericht erhoben werden, was im Kontext von Schule und Unterricht unüblich und darüber hinaus nicht unproblematisch ist. Eindrücke und Urteile hinsichtlich kognitiver Fähigkeiten von Lernenden können hingegen einfacher während des laufenden Schuljahres anhand regulärer Tests seitens der Lehrkräfte validiert und ggf. revidiert werden. (Autorinnen und Autoren, anonymisiert; Autorinnen und Autoren, anonymisiert).

Lehramtsstudierende überschätzten das FSK der Lernenden signifikant. Überschätzungen wurden bislang vor allem in Studien gefunden, die sich auf die Leistung der Lernenden bezogen (Bates und Nettelbeck, 2001; Urhahne et al., 2010). Hinsichtlich des FSK von Lernenden hat sich bislang jedoch gezeigt, dass dieses von Lehrkräften tendenziell eher unterschätzt wurde (Praetorius et al. 2011; Praetorius et al. 2010). Die anderen Merkmale (AB, WS) wurden hingegen relativ genau eingeschätzt. Eine Erklärung für die Überschätzung des FSK könnte sein, dass in die Beurteilung des FSK auch die Leistungen der Schülerinnen und Schüler einfließen. Wie Forschungsergebnisse zeigen, werden Leistungen meist überschätzt, unter anderem deshalb, weil für ihre Einschätzung oft auch andere Schülermerkmale herangezogen werden (Kaiser et al., 2012; Autorinnen und Autoren, anonymisiert): Nimmt die urteilende Person die Leistung als hoch wahr, so könnte sie dem zu beurteilenden Schüler auch höhere Werte im FSK zuschreiben.

In Bezug auf die Differenzierungskomponente zeigte sich, dass Lehramtsstudierende die Streuung aller Merkmale in der Gesamtstichprobe der Mädchen und Jungen signifikant überschätzten. Dieses Ergebnis deckt sich zumindest bezüglich des FSK mit vorherigen Arbeiten, in denen sich ebenfalls eine Überschätzung der Streuung des FSK gezeigt hat, wie dies oft auch bei anderen Merkmalen wie der Schülerleistung oder der Schülermotivation der Fall ist. Was die bislang bei AB und WS nicht untersuchte Differenzierungskomponente anbetrifft, erweitert die vorliegende Studie somit den bestehenden Kenntnisstand (z.B. Autorinnen und Autoren, anonymisiert; Spinath, 2005; Urhahne et al., 2010). Die für alle drei Konstrukte festgestellte Überschätzung der Streuung unterschied sich jedoch nicht signifikant zwischen Mädchen und Jungen. Stereotype Erwartungen, die zu einer Homogenisierung der Einschätzungen in der Mädchengruppe führen sollten, scheinen somit keine bedeutsame Rolle zu spielen und die angenommene Homogenisierungstendenz konnte nicht bestätigt werden.

Statistisch bedeutsame Unterschiede in der Urteilsakkuratheit (Rangkomponente) zwischen Mädchen und Jungen ließen sich bei den Lehramtsstudierenden lediglich für das FSK nachweisen: Unterschiede im FSK wurden bei Jungen deutlich besser erkannt als bei Mädchen. Dies könnte

anders als bei der Differenzierungskomponente für eine durch Stereotypenbildung begünstigte Homogenisierungstendenz bei den Mädchen oder eine androzentrische Sichtweise sprechen, bei der Jungen im Fach Physik stärkere Aufmerksamkeit geschenkt wird. Das Ergebnis könnte aber auch an in der Experimentalsituation begründet sein, in der Unterschiede im FSK bei den Jungen möglicherweise deutlicher sichtbar sind. Die Experimentalsituation ist eine neue, vergleichsweise unbekannte Situation mit einer unbekannten Lehrkraft. Es könnte sein, dass sich in einer solchen Situation Jungen mit einem günstigen Selbstkonzept besonders stark engagieren und verstärkt urteilsrelevante Cues produzieren, die dann zu genaueren Urteilen geführt haben könnten (Funder, 1995).

Die für Mädchen und Jungen signifikant unterschiedlichen Verschätzungstendenzen hinsichtlich der Niveaukomponente sind vereinbar mit der Annahme, dass das Unterrichtsfach Physik als genuin männlich wahrgenommen wird (Willems, 2007) und Jungen im Vergleich zu Mädchen mehr Fähigkeiten in Physik zugeschrieben werden (Kessels, 2008; Lightbody, et al., 1996; Solga und Pfahl, 2009; Tiedemann, 2000). Analog verhält es sich mit den differenziellen Verschätzungen der WS, welche ebenfalls auf stereotype Erwartungen hindeuten. In der Wahrnehmung der Lehramtsstudierenden wurde die Schwierigkeit der Physikaufgaben von den Jungen unterschätzt, von den Mädchen dagegen überschätzt.

Die Einschätzungen hinsichtlich der Streuung der nicht-kognitiven Merkmale haben ergeben, dass Lehramtsstudierende diese in Bezug auf die WS und AB bei den Jungen und die WS bei den Mädchen signifikant überschätzten, sich die Einschätzungen jedoch nicht signifikant voneinander unterschieden. Somit wurden die Lerngruppen hinsichtlich dieser Merkmale als inhomogener wahrgenommen, als sie tatsächlich waren.

5 Studie 2

In der zweiten Studie wurden parallel zur ersten dieselben Forschungsfragen überprüft. Unter Beibehaltung der Hypothesen, wurden in dieser Studie allerdings erfahrene Lehrkräfte befragt.

5.1 Methode

In der zweiten Studie wurden die gleichen Messinstrumente verwendet wie in der ersten Studie. Ebenso sind die statistischen Analysen gleichbleibend. Daher werden an dieser Stelle nur die Unterschiede zur ersten Studie hervorgehoben.

5.2 Stichprobe und Design

An der zweiten Studie nahmen N = 10 erfahrene Lehrkräfte (50% weiblich, M = 37.9 Jahre, SD = 8.38) als Urteilende teil. Als Referenzwerte für die Urteile wurden auch in dieser Studie die Selbstberichtsdaten der N = 207 Lernenden (40.3 % weiblich, M = 15.6 Jahre, SD = 0.63) herangezogen. Die Lernenden von zwei der Schulklassen besuchten zum Zeitpunkt der Studiendurchführung die Realschule, die der restlichen acht Schulklassen das Gymnasium. Bei den urteilenden Lehrkräften handelte es sich um jene Personen, die regulär den Physikunterricht der Lernenden in der Studie leiteten. Die Klassenstärke betrug im Mittel 21 Schüler (min. = 17; max. = 23).

5.3 Ergebnisse Studie 2

5.3.1 Voranalysen und Deskriptive Statistiken

Die Voranalysen und Deskriptiva, die sich auf die Schülerschaft beziehen, können dem entsprechenden Ergebnisabschnitt bei Studie eins sowie der Tab. 1 entnommen werden. Hinsichtlich der Interkorrelationen in Tab. 2 wird ersichtlich, dass die Lehrkrafturteile der jeweiligen Schülermerkmale jeweils stärker miteinander zusammenhingen als die Schülerangaben untereinander.

5.3.2 Genauigkeit der Lehrkrafteinschätzungen

In Tab. 4 sind die Ergebnisse zur Urteilsgenauigkeit der Lehrkräfte dargestellt. In Bezug auf die erste Fragestellung, wie Lehrkräfte die verschiedenen Schülermerkmale einschätzen, zeigte sich, dass die mittleren Rangkomponenten für das FSK (r = .54, $t(9)$ = 11.32, p < .01) und die AB (r = .30, $t(9)$ = 2.85, p < 0.5) mäßig bis mittelhoch ausfielen und signifikant von 0 abwichen.

Tabelle 4

Rang-, Niveau- und Differenzierungskomponente Studie 2 (Lehrkräfte) für die Gesamtstichprobe sowie Substichproben der Mädchen und Jungen

	Gesamt		Mädchen		Jungen		Mädchen vs. Jungen
	M	SD	M	SD	M	SD	M_{diff}
Rangkomponente							
FSK	.54 **	0.16	.45*	0.31	.62**	0.15	-.17+
AB	.30*	0.29	.00	0.45	.21	0.33	-.21
WS	.15+	0.24	.11	0.37	.15	0.22	-.04
Niveaukomponente							
FSK	0.23 *	0.26	0.13	0.36	0.29*	0.27	-0.16
AB	-0.21	0.41	-0.13	0.54	-0.19	0.45	-0.06
WS	-0.13	0.34	-0.04	0.57	-0.23	0.43	-0.19
Differenzierungs-komponente							
FSK	1.30 *	0.26	1.38*	0.40	1.27+	0.39	0.11
AB	1.62 *	0.54	1.46*	0.71	1.70*	0.18	0.24
WS	1.83*	0.39	0.93+	0.31	1.76*	0.62	0.83**

Anmerkungen. Testungen gegen 0 bzw. 1 mittels *t*-Test;
Vergleich Mädchen vs. Jungen: Wilcoxon-Test; + $p < .10$; * $p < .05$, ** $p <. 01$

Die mittlere Rangkomponente der WS war dagegen recht niedrig und wich nur marginal signifikant von 0 ab ($r = .15$, $t(9) = 1.85$, $p < .10$). Auch bei den Lehrkräften bestanden starke interindividuelle Unterschiede in der Urteilsakkuratheit ($r = -.32 \leq r \leq .84$). Wie der Tab. 4 weiter zu entnehmen ist, überschätzten die Lehrkräfte das FSK signifikant ($M = 0.23$, $t(9) = 2.76$, $p < 0.05$). Des Weiteren zeigte sich, dass Lehrkräfte die Streuung aller Merkmale signifikant überschätzten: FSK ($M = 1.30$, $t(27) = 9.93$, $p < .05$), WS ($M = 1.83$, $t(9) = 6.79$, $p < .05$) sowie AB ($M = 1.62$, $t(9) = 3.64$, $p < .05$). Die Daten unterstützen die Hypothese 1 daher nur in Teilen.

5.3.3 Unterschiede in der Genauigkeit der Einschätzung von Merkmalsniveau und Merkmalsunterschieden bei Jungen und Mädchen

Im Hinblick auf die zweite Fragestellung wurde zunächst wieder für Mädchen und Jungen separat geprüft, ob sich die Mittelwerte der Diagnosekennwerte in beiden Gruppen signifikant von Null bzw. im Fall der Differenzierungskomponente von Eins unterscheiden. Im zweiten Schritt wurde dann ebenfalls wieder geprüft, ob sich für die drei Diagnosekennwerte signifikante Unterschiede zwischen Mädchen und Jungen nachweisen lassen. Für die Rangkomponente zeigte sich, dass deren Mittelwert für das FSK sowohl bei den Jungen ($M = .62$, $\underline{t}(9) = 13.04$, $p < .01$) als auch bei den Mädchen ($M = .45$ $\underline{t}(9) = 3.93$, $p < .05$) signifikant von 0 abwich, während die übrigen Werte niedrig und nicht signifikant waren. Was die Unterschiede zwischen Mädchen und Jungen betrifft, zeigte sich, dass lediglich der Unterschied beim FSK (M (Mädchen) = .45 vs. M (Jungen), $z = -1.79$, $p < .10$) marginal signifikant zugunsten der Jungen ausfiel. Die Effektstärke (berechnet anhand z geteilt durch die Wurzel aus n) liegt bei $r = .57$ und entspricht nach Cohen (1988) einem starken Effekt, der angesichts der geringen Stichprobe trotzdem nur marginal signifikant ausfällt. Die Unterschiede zwischen den mittleren Rangkomponenten der Jungen und Mädchen bei den Merkmalen AB und WS waren hingegen nicht statistisch signifikant. Hinsichtlich der Rangkomponente wurde Hypothese 2 also nur in Teilen durch die Daten gestützt.

Bei der separaten Berechnung der Niveaukomponente für Mädchen und Jungen zeigte sich, dass Mädchen bei allen drei Merkmalen weder signifikant über- noch unterschätzt wurden. Bei den Jungen wurde hingegen das FSK signifikant überschätzt ($M = .29$, $t(9) = 3.36$, $p < .05$) während bei den anderen Merkmalen keine signifikante Über- oder Unterschätzungen zu verzeichnen waren; vgl. Tab. 4). Außerdem gab es bei allen drei Merkmalen keine signifikanten Unterschiede zwischen Mädchen und Jungen. In Bezug auf die Niveaukomponente unterstützten die Daten die Hypothese 2 somit nicht.

Im Falle der Differenzierungskomponente zeigte sich, dass die Lehrkräfte bei den Mädchen die Streuung für das FSK ($M = 1.38$, $t(9) = 2.97$, $p < .05$) und AB ($M = 1{,}46$, $t(9) = 2.04$, $p < .05$) signifikant überschätzten, wohingegen die WS tendenziell unterschätzt wurde ($M = 0.93$, $t(9) = -0.74$, $p < .10$; vgl. Tab. 4). Bei den Jungen zeichnete sich ein leicht anderes Bild ab. Es ergab sich, dass Lehrkräfte die Streuung der Merkmale AB ($M = 1.70$, $t(9) = 3.91$, $p < .05$) und WS ($M = 1{,}76$, $t(9) = 3.88$, $p < .05$) signifikant und die Streuung des FSK marginal signifikant überschätzten ($M = 1{,}27$, $t(9) = 2.19$, $p < .10$; vgl. Tab. 4). Ein signifikanter Unterschied in der

Differenzierungskomponente zwischen Mädchen und Jungen trat lediglich bei der WS (M (Mädchen) = 0.93, z = -2.60, $p < .01$) auf, während sich für die Merkmale FSK und AB keine signifikanten Geschlechtsunterschiede ergaben. Für die Variable WS wurde Hypothese 2 also von den Daten gestützt.

5.4 Ergebnisdiskussion Studie 2

Im Fokus der zweiten Studie stand die Urteilsakkuratheit von Lehrkräften im Fach Physik. Parallel zu Studie 1 wurde auf die weniger stark beforschten Schülercharakteristika physikbezogenes FSK, WS und AB fokussiert. Auch in dieser Arbeit wurde die Urteilsakkuratheit sowohl allgemein, als auch nach Geschlecht der beurteilten Lernenden aufgeschlüsselt, analysiert. Die Ergebnisse der zweiten Studie sind denen der ersten Studie ähnlich. Demgemäß sind die Ergebnisse im Bezug auf den Forschungskontext ähnlich zu verorten und zu interpretieren. Um Doppelungen zu vermeiden, wird an dieser Stelle daher nur auf Ergebnisse eingegangen, bei denen eine ergänzende respektive andere Einordnung vorzunehmen ist.

Lehrkräfte schätzten die Niveaukomponente für Mädchen und Jungen nicht signifikant unterschiedlich ein, so dass nicht davon ausgegangen werden kann, dass ein Gender Bias vorlag. Dies steht in Einklang mit einer Arbeit von Hofer (2015). In dieser konnte gezeigt werden, dass sich der Gender Bias im Unterrichtsfach Physik mit zunehmender Lehrerfahrung abzumildern scheint. Bei den Ergebnissen zur Differenzierungskomponente zeigte sich, dass lediglich im Falle der WS die Merkmalsstreuung für die Mädchen signifikant niedriger eingeschätzt wurde als bei den Jungen, was darauf hindeutet, dass die Gruppe der Mädchen durch die Lehrkräfte als homogener wahrgenommen wurde. Dieses Ergebnis kann als ein erster Hinweis darauf gedeutet werden, dass systematische, auf das Geschlecht der Schülerinnen und Schüler zurückführbare Verzerrungstendenzen auch bei erfahrenen Lehrkräften eine Rolle für die Urteilsgenauigkeit zu spielen scheinen. Allerdings darf das Ergebnismuster für die Niveau- und Differenzierungskomponente angesichts der kleinen Stichprobe nicht überinterpretiert werden.

6 Abschließende Gesamtdiskussion

Aufgrund der Bedeutung der Untersuchung der Urteilsakkuratheit und des Zustandekommens von interindividuellen Unterschieden in der Akkuratheit zwischen Urteilenden,

wurden im Rahmen dieser Arbeit zwei Studien durchgeführt. Studie 1 nahm Lehramtsstudierende in den Blick, wohingegen Studie 2 auf Lehrkräfte fokussierte. In beiden Studien wurde die Urteilsakkuratheit hinsichtlich leistungsbezogener nicht-kognitiver Schülermerkmale im Fach Physik, einer in diesem Forschungskontext bislang selten untersuchten Domäne, untersucht. Im Vordergrund stand dabei der Gender Bias als eine Erklärungsmöglichkeit für individuelle Unterschiede in der Urteilsgenauigkeit. Beide Studien unterschieden sich in verschiedener Hinsicht, so dass ein direkter Vergleich nur sehr bedingt möglich ist. Fasst man die Ergebnisse beider Studien zusammen, so zeigt sich in der Zusammenschau aber ein einigermaßen robustes Bild.

Die Rangkomponente wurde von beiden Urteilergruppen für das FSK mäßig genau eingeschätzt, für AB und WS dagegen weniger genau. Möglicherweise gibt es für das breiter angelegte Merkmal FSK mehr und deutlichere Hinweisreize (cues) als für die recht spezifischen Merkmale AB und WS. Beim FSK wurden – von den Lehramtsstudierenden recht deutlich, von den erfahrenen Lehrkräften ansatzweise – Jungen genauer beurteilt als Mädchen. Dies scheint für die Sichtweise von Physik als typisch männlichem Fach zu sprechen, bei dem Jungen eher im Fokus der Aufmerksamkeit stehen als Mädchen, deren Leistungen in diesem Fach weniger beachtet werden und vielleicht auch weniger sichtbar sind. Dafür spricht auch das höhere FSK der Jungen, die gleichzeitig auch dazu neigen, Aufgabenstellungen als leichter wahrzunehmen und weniger Anstrengung investieren.

Das FSK der Lernenden wird von beiden Urteilergruppen, von den Lehramtsstudierenden stärker als von erfahrenen Lehrkräften, überschätzt. Bei den Lehramtsstudierenden wird zudem das FSK der Jungen deutlich überschätzt und signifikant höher eingeschätzt als das der Mädchen. Das ist tendenziell auch bei den Lehrkräften der Fall, aber deutlich schwächer und nicht signifikant. Bei den Lehrkräften gibt es für AB und WS keine signifikanten Fehleinschätzungen im Niveau, während Lehramtsstudierendendie die AB der Jungen und die WS der Mädchen überschätzen. Die geringe Zahl signifikanter Fehleinschätzungen im Niveau von AS und WS kann nicht ohne weiteres als Hinweis auf eine hohe Niveaugenauigkeit angesehen werden, zumal das Urteilskriterium hier keine objektive Testleistung ist, sondern eine Selbsteinschätzung der Beurteilten. Generell kann das Fehlen von signifikanten Abweichungen von der Nullhypothese nicht als Bestätigung für deren Richtigkeit aufgefasst werden (siehe etwa Bortz und Döring, 2006).

Möglicherweise haben sich im vorliegenden Fall Urteiler und Beurteilte einfach gleichermaßen stark an der Mittelkategorien der Urteilsskala als Anker orientiert.
Bei der Differenzierungskomponente zeigt sich ausnahmslos eine deutliche Überschätzung der Merkmalsstreuung, was auf eine allgemeine Neigung, die extremen Ausprägungen zu akzentuieren, hindeuten könnte. Möglicherweise spielen hier kognitive Vereinfachungstendenzen eine Rolle, bei denen sich die Urteilenden an zwei Stereotypen, dem des guten und des schlechten Schülers orientieren. Die erwartete Homogenisierungstendenz bei der Mädchengruppe zeigt sich lediglich im Falle der erfahrenen Lehrkräfte bei WS.

Insgesamt zeigt sich ein Gender Stereotyp zwar ansatzweise, aber nicht durchgängig und wenn, dann eher bei den Lehramtsstudierenden als bei den erfahrenen Lehrkräften. Auf den ersten Blick hat es den Anschein, also könnten Gender Stereotype mit fehlender oder geringer Berufserfahrung zursammenhängen. Angesichts der mangelnden Vergleichbarkeit beider Urteilergruppen kann dies aber nicht zwingend auf Unterschiede zwischen den Urteilenden zurückgeführt werden. Im Folgenden werden Stärken wie Limitationen der Arbeit sowie Implikationen für Forschung und Praxis diskutiert.

6.1 Stärken und Limitationen der Arbeiten

Zu den Stärken der vorliegenden Arbeit gehört, dass der Frage nach möglichen kognitiven Verzerrungen bei der Urteilsbildung am Beispiel des Gender Bias im bisher eher selten untersuchten Unterrichtsfach Physik nachgegangen wurde, und dabei die nicht-kognitiven Merkmale WS und AB, die im Kontext diagnostischer Kompetenz bislang kaum untersucht wurden, zusammen mit dem FSK der Lernenden in den Blick genommen wurden. WS und AB wurden dabei mehrfach im Verlauf einer realen Unterrichtssituation erhoben, wodurch eine - im Vergleich zur Einfachmessung - fundiertere Erhebung des Konstrukts bei den Schülern sichergestellt werden konnte. Als eine weitere Stärke der vorliegenden Arbeit kann auch der Einbezug von Lehramtsstudierenden angesehen werden, weil dies erste Hinweise auf Entstehung oder Reduktion von stereotypen Erwartungen geben könnte. Insofern erweitert die vorliegende Arbeit den vorliegenden Kenntnisstand in verschiedener Hinsicht.

Limitationen dieser Arbeit sind vorwiegend methodischer Natur. Es handelt sich um eine relativ kleine Lehramtsstudierenden- (Studie 1) und Lehrkraftstichprobe (Studie 2). Offen bleibt weiterhin, in welcher Weise sich die Experimentiersituation, in der der Unterricht der

Lehramtsstudierenden stattfand, auf deren Urteilsbildung auswirkte. Des Weiteren ist unklar, wie relevant bzw. bekannt die Indikatoren waren, die den Urteilern für ihre Urteile zur Verfügung standen und wie sichtbar die Merkmale der Lernenden für die Urteilenden waren (Funder, 2012).

6.2 Implikationen für Forschung und Praxis

Ausgehend von den berichteten Ergebnissen und Limitationen der Arbeit ergeben sich verschiedene Forschungsdesiderata. Zunächst müssten die berichteten Ergebnisse anhand einer größeren Stichprobe repliziert werden Insbesondere wäre eine bessere Vergleichbarkeit der Urteilssituation zu gewährleisten: neben einer gleichen Anzahl der Beurteilten vor allem eine vergleichbare Vertrautheit beider Urteilergruppen mit den Lernenden. Denkbar wäre, dass auch Lehrkräfte ihnen unbekannte Schülerinnen und Schüler beurteilen könnten. Damit könnten auch Bezugssysteme, die sich durch die Erfahrung mit einem Klassenverband herausbilden und die Urteilsbildung von Lehrkräften maßgeblich beeinflussen, wegfallen.

Die berichteten Ergebnisse deuten darauf hin, dass sich ein Gender Bias vor allem bei Berufsanfängern zeigt und sich mit zunehmender Berufserfahrung abschwächt. Eine sich unmittelbar anschließende Forschungsfrage beträfe demnach die Rolle der Berufserfahrung für den Gender Bias. So könnte im Rahmen von Längsschnittstudien analysiert werden, welche Rolle die Länge der vorangegangenen Interaktionen mit den Lernenden sowohl für die Urteilsakkuratheit als auch die Ausprägung des Gender Bias bei den Urteilen spielt. In diesem Zusammenhang kann auf den sogenannten Zero-Acquaintance Ansatz (vgl. Ambady, Hallahan und Rosenthal, 1995) zurückgegriffen werden, der in der Urteilsforschung im Bereich der Persönlichkeits- und Sozialpsychologie häufig herangezogen wird, um die Genese von Persönlichkeitsurteilen eingehend und unabhängig von den Einflüssen vorangegangener Interaktionen zu untersuchen. Auch in der Lehrkrafturteilsforschung gibt es erste Studien, die anhand dieses Ansatzes die Rolle der Dauer der Interaktion zwischen Lehrkräften und Lernenden für die Urteilsakkuratheit untersucht haben (z.B. Praetorius et al. 2014). Weiterhin könnte im Rahmen einer quasi-experimentellen Studie der Frage nachgegangen werden, ob sich der Gender Bias bei den Urteilen durch eine Sensibilisierung von Lehramtsstudierenden bereits in der Ausbildungsphase positiv beeinflussen lässt.

Wenn sich die Ergebnisse zum Gender Bias bestätigen lassen, ließe sich daraus für die Praxis die Forderung ableiten, bereits im Lehramtsstudium, aber auch in der späteren Schulpraxis

für diese Thematik zu sensibilisieren. Zu diesem Zweck könnte über die Entwicklung und statistische Evaluation von Interventionen nachgedacht werden, in denen Physik als Fach dargestellt wird, in dem Leistungen weniger auf (männlichen) Fähigkeiten, als auf der individuellen AB beruhen (Kessels, 2015). Eine genauere Einschätzung und Wahrnehmung des FSK, WS und AB durch (angehende) Lehrkräfte und das Verständnis, dass eine verfälschte Wahrnehmung der Fähigkeiten, welche sich auf die Benotung auswirkt, auch Folgen für Mädchen in Bezug auf naturwissenschaftliche Fächer haben kann, wäre hierbei wünschenswert. Akkuratere resp. leicht positiv verzerrte Einschätzungen und Rückmeldungen könnten sich entsprechend positiv auf das FSK, welches in Zusammenhang mit der Leistung steht, auswirken. Dies wiederum könnte zur Folge haben, dass Mädchen stärker für den MINT Bereich begeistert werden und ein stärkeres Interesse am Fach ausbilden (z.B. Häußler und Hoffmann, 1995).

Eine Sensibilisierung von Lehrkräften ist auch deshalb wünschenswert, weil sich ein Gender Bias möglicherweise im Sinne einer sich selbsterfüllenden Prophezeiung (Rosenthal und Jacobson, 1968) auf die Lernenden auswirken könnte. Durch differenzielle instruktionale Aufgaben- oder Hilfestellungen sowie Anspruchshaltungen (z.B. Woodcock und Vialle, 2011), die unterschiedliche Attribuierung von Leistung und daraus resultierende Effekte auf schülerische Selbstwirksamkeiterwartungen (z. B. Weiner, 2000) oder die unterschiedliche Bewertung objektiv gleicher Produkte und Leistungen (Hofer, 2015) können Lernende somit auf implizite oder explizite Art und Weise in ihrem Lernverhalten beeinflusst werden. Dies bewirkt - bzw. verstärkt - möglicherweise in der Folge die in Forschung und Unterricht wiederholt gefundenen Unterschiede in Leistung und FSK von Jungen und Mädchen (Hofer und Stern, 2016; Kessel et al., 2008; Reiss et al, 2016), insbesondere in den Naturwissenschaften.

Da das unterrichtliche Handeln von Lehrkräften als ein wichtiger Faktor für die Ausbildung positiver Einstellungen von Lernenden in Bezug auf naturwissenschaftliche Fächer ausgemacht wurde (für einen Überblick: Osborne, Simon, und Collins, 2003), könnten akkuratere Lehrkrafturteile im Kontext des Physikunterrichts Möglichkeiten zur Modifikation negativer Einstellungen von Lernenden über entsprechend adaptive Lehr-Lernsettings bieten und in einer langzeitlichen Perspektive auch zu einem höheren Anteil an Mädchen in MINT-Fächern und darüber hinausgehend in MINT-Berufen führen.

Literaturverzeichnis

Autorinnen & Autoren (anonymisiert)

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

Ambady, N., Hallahan, M., & Rosenthal, R. (1995). On judging and being judged accurately in zero-acquaintance situations. *Journal of Personality and Social Psychology*, 69(3), 518–529. https://doi.org/10.1037/0022-3514.69.3.518

Anders, Y., Brunner, M., & Krauss, K. (2011). Diagnostic skills of mathematics teachers and the performance of their students. *Psychology in Education*, 3(2), 175–193. https://doi.org/10.1007/978-3-319-66327-2_2

Anders, Y., Kunter, M., Brunner, M. Krauss, S., & Baumert, J. (2010). Diagnostische Fähigkeiten von Mathematiklehrkräften und ihre Auswirkungen auf die Leistungen ihrer Schülerinnen und Schüler. *Psychologie in Erziehung und Unterricht*, 57, 175–193. https://doi.org/10.2378peu2010.art13d

Artelt, C., Stanat, P., Schneider, W., & Schiefele, U. (2001). Lesekompetenz: Testkonzeption und Ergebnisse. In J. Baumert, E. Klieme, M. Neubrand, M. Prenzel, U. Schiefele, W. Schneider, P., Stanat, K.-J. Tillmann, & M. Weiß (Hrsg.), *PISA 2000: Basiskompetenzen von Schülerinnen und Schülern im internationalen Vergleich* (S. 69–137). Opladen: Leske + Budrich.

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. https://doi.org/10.1207/s15326977ea1103&4_2

Baumert, J., & Kunter, M. (2006). Stichwort: Professionelle Kompetenz von Lehrkräften. *Zeitschrift für Erziehungswissenschaft*, 9, 469–520. https://doi.org/10.1007/s11618-006-0165-2

Baumert, J. & Lehmann, R. (1997). *TIMSS - Mathematisch-naturwissenschaftlicher Unterricht im internationalen Vergleich - Deskriptive Befunde.* Wiesbaden: VS Verlag für Sozialwissenschaften.

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

Beck, E., Brühwiler, C., & Müller, P. (2007). Adaptive Lehrkompetenz als Voraussetzung für individualisiertes Lernen in der Schule. In D. Lemmermöhle, M. Rothgangel, S. Bögeholz, M. Hasselhorn, & R. Watermann (Hrsg.), *Professionell lehren - Erfolgreich lernen* (S. 197–210). Münster: Waxmann.

Behrmann L., & Souvignier, E. (2013). The relation between teachers' diagnostic sensitivity, their instructional activities, and their students' achievement gains in reading. *Zeitschrift für Pädagogische Psychologie*, 27(4), 2013, 283–293. https://doi.org/10.1024/1010-0652/a000112

Bortz, J. & Döring, N. (2006). *Forschungsmethoden und Evaluation für Human- und Sozialwissenschaftler* (4. Aufl.). Heidelberg: Springer.

Cohen, J. (1988). *Statistical power analysis for the behavioral sciences*. New York, NY: Routledge Academic.

Corno, L. (2008). On teaching adaptively. *Educational Psychologist*, 43(3), 161–173. https://doi.org/10.1080/00461520802178466

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

Ditton, H., & Krüsken, J. (2006). Der Übergang von der Grundschule in die Sekundarstufe I. *Zeitschrift für Erziehungswissenschaft*, 9(3), 348–372. https://doi.org/10.1007/s11618-006-0055-7

Dovidio, J. F., Hewstone, M., Glick, P., & Esses, V.M. (2010). Prejudice, stereotyping and discrimination: Theoretical and empirical overview. In J. F. Dovidio, M. Hewstone, P. Glick & V. M. Esses (Hrsg.). *The Sage Handbook of Prejudice, Stereotyping and Discrimination*. London: Sage Publications.

Efklides, A., & Tsiora, A. (2002). Metacognitive experiences, self-concept, and self-regulation. *Psychologia: An International Journal of Psychology in the Orient*, 45(4), 222–236. https://doi.org/10.2117/psysoc.2002.222

Eichler, M., Fuchs, J., & Maschewsky-Schneider, U. (2000). Richtlinien zur Vermeidung von Gender Bias in der Gesundheitsforschung. *Zeitschrift für Gesundheitswissenschaften*, 8(4), 293. https://doi.org/10.1007/BF02955909

Ellemers, N. (2018). Gender stereotypes. Annual Review of Psychology, 69, 275–298. https://doi.org/10.1146/annurev-psych-122216-011719

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. https://doi.org/10.3200/JOER.102.6.453-462

Funder, D. C. (1995). On the accuracy of personality judgment: A realistic approach. *Psychological Review*, 102, 652–670. https://doi.org/10.1037/0033-295x.102.4.652

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

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

Hascher, T. (2003). Diagnose als Voraussetzung für gelingende Lernprozesse. *Journal für Lehrerinnen- und Lehrerbildung*, 2, 25–30.

Hascher, T. (2008). Diagnostische Kompetenzen im Lehrerberuf. In C. Kraler & M. Schratz (Hrsg.), *Wissen erwerben, Kompetenzen entwickeln. Modelle zur kompetenzorientierten Lehrerbildung* (S. 71–86). Münster: Waxmann.

Häußler, P., & Hoffmann, L. (1995). Physikunterricht - an den Interessen von Mädchen und Jungen orientiert. *Unterrichtswissenschaft*, 23, 107–126.

Helmke, A. (2009). Unterrichtsqualität und Lehrerprofessionalität. Seelze-Velber: Klett-Kallmeyer. Herwartz-Emden, L., Schurt, V., & Waburg, W. (2012). *Mädchen und Jungen in Schule und Unterricht.* Stuttgart: Kohlhammer.

Hofer, S. I. (2015). Studying gender bias in physics grading: The role of teaching experience and country. *International Journal of Science Education*, 37, 2879-2905. https://doi.org/10.1080/09500693.2015.1114190

Hofer, S. I., & Stern, E. (2016). Underachievement in physics: When intelligent girls fail. *Learning and Individual Differences*, 51, 119-131. https://doi.org/10.1016/j.lindif.2016.08.006

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

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

Jansen, M., Schroeders, U., Lüdtke, O., & Marsh, H. W. (2019). The dimensional structure of students' self-concept and interest in science depends in course composition. *Learning and Instruction*, 60, 20-28. https://doi.org/10.1016/j.learninstruc.2018.11.001

Jussim, L., & Eccles, J. S. (1992). Teacher expectations II: Construction and reflection of student achievement. *Journal of Personality and Social Psychology*, 63, 947–961. https://doi.org/10.1037/0022-3514.63.6.947

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

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? *Journal for Educational Research Online*, 3, 119–147.

Karst, K., Schoreit, E., & Lipowsky, F. (2014). Diagnostische Kompetenzen von Mathematiklehrern und ihr Vorhersagewert für die Lernentwicklung von Grundschulkindern. *Zeitschrift für Pädagogische Psychologie*, 28(4), 237–248. doi: 10.1024/1010-0652/a000133

Kessels, U. (2008). Physikinteresse gilt als unweiblich: Erklärung für die Unterrepräsentanz von Mädchen und Frauen im MINT-Bereich. *Schule im Blickpunkt*, 2, 24–26.

Kessels, U., & Hannover, B. (2004). Empfundene „Selbstnähe" als Mediator zwischen Fähigkeitsselbstkonzept und Leistungskurswahlintentionen. *Zeitschrift für Entwicklungspsychologie und Pädagogische Psychologie*, 36(3), 130–138. https://doi.org/10.1026/0049-8637.36.3.130

Kessels, U., Rau, M., & Hannover, B. (2006). What goes well with physics? Measuring and altering the image of science. *British Journal of Educational Psychology*, 74, 761–780.

https://doi.org/10.1348/000709905X59961

Kessels, U. (2015). Bridging the gap by enhancing the fit: How stereotypes about STEM clash with stereotypes about girls. *International Journal of Gender, Science and Technology*, 7(2), 280-296.

Kessels, U., Warner, L. M., Holle, J., & Hannover, B. (2008). Identitätsbedrohung durch positives schulisches Leistungsfeedback - Die Erledigung von Entwicklungsaufgaben im Konflikt mit schulischem Engagement. *Zeitschrift für Entwicklungspsychologie und Pädagogische Psychologie*, 40(1), 22–31. https://doi.org/10.1026/0049-8637.40.1.22

Khalaila, R. (2014). The relationship between academic self-concept, intrinsic motivation, test anxiety, and academic achievement among nursing students: Mediating and moderating effects. *Nurse Education Today*, 35, 432–438. https://doi.org/10.1016/j.nedt.2014.11.001

KMK (2019). Standards für die Lehrerbildung: Bildungswissenschaften (Beschluss der Kultusministerkonferenz vom 16.12.2004 i. d. F. vom 16.05.2019) Abgerufen am 29.09.2020 von https://www.kmk.org/themen/allgemeinbildende-schulen/lehrkraefte/lehrerbildung.html

Krauss, S., Neubrand, M., Blum, W., Baumert, J., Brunner, M., Kunter, M., & Jordan, A. (2008). Die Untersuchung des professionellen Wissens deutscher Mathematik-Lehrerinnen und -Lehrer im Rahmen der COACTIV-Studie. *Journal für Mathematik-Didaktik*, 29(3-4), 233–258. https://doi.org/10.1007/BF03339063

Krolak-Schwerdt, S., Böhmer, M., & Gräsel, C. (2009). Verarbeitung von schülerbezogener Information als zielgeleiteter Prozess: Der Lehrer als «flexibler Denker». *Zeitschrift für Pädagogische Psychologie*, 23, 175–186. https://doi.org/10.1024/1010-0652.23.34.175

Kuhl, P. & Hannover, B. (2012). Differenzielle Benotungen von Mädchen und Jungen. *Zeitschrift für Entwicklungspsychologie und Pädagogische Psychologie*, 44(3), 153-162. https://doi.org/10.1026/0049-8637/a000066

Leslie, S. J., Cimpian, A., Meyer, M., & Freeland, E. (2015). Expectations of brilliance underlie gender distributions across academic disciplines. *Science,* 347(6219), 262-265. https://doi.org/10.1126/science.1261375

Lightbody, P., Siann, G., Stocks, R., & Walsh, D. (1996). Motivation and attribution at secondary school: The role of gender. *Educational Studies* 22, 13–25. https://doi.org/10.1080/0305569960220102

Lipnevich, A. A., & Roberts, R. D. (2012). Noncognitive skills in education: Emerging research and applications in a variety of international contexts. *Learning and Individual Differences*, 22, 173–177. https://doi.org/10.1016/j.lindif.2011.11.016

Lorenz, C., & Artelt, C. (2009). Fachspezifität und Stabilität diagnostischer Kompetenz von Grundschullehrkräften in den Fächern Deutsch und Mathematik. *Zeitschrift für Pädagogische Psychologie*, 23, 211–222. https://doi.org/10.1024/1010-0652.23.34.211

Maaz, K., Neumann, M., Trautwein, U., Wendt, W., Lehmann, R., & Baumert, J. (2008). Der Übergang von der Grundschule in die weiterführende Schule. Die Rolle von Schüler- und Klassenmerkmalen beim Einschätzen der individuellen Lernkompetenz durch die Lehrkräfte. *Schweizerische Zeitschrift für Bildungswissenschaften*, 30(3), 519–548.

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

Marsh, H. W., & Craven, R. (1997). Academic self-concept: Beyond the dustbowl. In G. Phye (Hrsg.), *Handbook of classroom assessment: Learning, achievement, and adjustment* (S. 131–198). Orlando, FL: Academic Press.

Marsh, H. W., Trautwein, U., Lüdtke, O., Köller, O., & Baumert, J. (2005). Academic self-concept, interest, grades, and standardized test scores: Reciprocal effects models of causal ordering. *Child Development*, 76(2), 397–416. https://doi.org/10.1111/j.1467-8624.2005.00853.x

Merzyn, G. (2008). *Naturwissenschaften, Mathematik, Technik – Immer unbeliebter?* Hohengehren: Schneider-Verlag GmbH.

Merzyn, G. (2009). Polarisierender Physikunterricht. *Physik in unserer Zeit*, 40, 312–313. https://doi.org/10.1002/piuz.200990112

Metallidou, P., & Efklides, A. (2001). The effects of general success-related beliefs and specific metacognitive experiences on causal attributions. In A. Efklides, J. Kuhl & R. M. Sorrentino (Hrsg.), *Trends and prospects in motivation research* (pp. 325- 347). Springer, Dordrecht.

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

Muckenfuß, H. (1995). *Lernen im sinnstiftenden Kontext: Entwurf einer zeitgemäßen Didaktik des Physikunterrichts.* Berlin: Cornelsen.

Osborne, J, Simon, S., & Collins, S. (2003) Attitudes towards science: A review of the literature and its implications, *International Journal of Science Education*, 25(9), 1049–1079. doi: 10.1080/0950069032000032199

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. https://doi.org/10.1177/0021934705285955

Pielmeier, M., Huber, S., & Seidel, T. (2018). Is teacher judgment accuracy of students' characteristics beneficial for verbal teacher-student interactions in classrooms? *Teaching and Teacher Education*, 76, 255–266. https://doi.org/10.1016/j.tate.2018.01.002

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

Pit-ten Cate, I. M, Krolak-Schwerdt, S., Glock. S. & Markova, M. (2014). Improving teachers' judgments: Obtaining change through cognitive processes. In S. Krolak-Schwerdt, S. Glock & M. Böhmer (Hrsg.), *Teachers' professional development: Assessment, training, and learning* (S. 45-62). Rotterdam: Sense Publisher.

Praetorius, A. K., Greb, K., Lipowsky, F., & Gollwitzer, M. (2010). Lehrkräfte als Diagnostiker. Welche Rolle spielt die Schülerleistung bei der Einschätzung von mathematischen Selbstkonzepten? *Journal for Educational Research Online*, 2, 121-144.

Praetorius, A.-K., Karst, K., Dickhäuser, O., & Lipowsky, F. (2011). Wie gut schätzen Lehrer die Fähigkeitsselbstkonzepte ihrer Schüler ein? Zur diagnostischen Kompetenz von Lehrkräften. *Psychologie in Erziehung und Unterricht*, 58, 81–91. https://doi.org/10.2378/peu2010.art30d

Praetorius, A.-K., Lipowsky, F., & Karst, K. (2012). Diagnostische Kompetenz von Lehrkräften: Aktueller Forschungsstand, unterrichtspraktische Umsetzbarkeit und

Bedeutung für den Unterricht. In R. Lazarides & A. Ittel (Hrsg.), *Differenzierung im mathematisch-naturwissenschaftlichen Unterricht* (S. 115–146). Bad Heilbrunn: Klinkhardt.

Praetorius, A.-K., Drexler, K., Rösch, L., Christophel, E., Heyne, N., Scheunpflug, A., Zeinz, H., & Dresel, M. (2015). Judging students' self-concepts within 30s? Investigating judgement accuracy in a zero-acquaintance situation. *Learning and Individual Differences*, 37, 231–236. https://doi.org/10.1016/j.lindif.2014.11.015

Praetorius, A.-K., & Südkamp, A. (2017). Eine Einführung in das Thema der diagnostischen Kompetenz von Lehrkräften. In A. Südkamp & A.-K. Praetorius (Hrsg.), *Diagnostische Kompetenz von Lehrkräften: Theoretische und methodische Weiterentwicklungen*. Münster: Waxmann.

Prenzel, M., Reiss, K., & Hasselhorn, M. (2009). Förderung der Kompetenzen von Kindern und Jugendlichen. In J. Milberg (Hrsg.), *Förderung des Nachwuchses in Technik und Naturwissenschaft* (S. 15–60). Heidelberg: Springer.

Reiss, K., Sälzer, C., Schiepe-Tiska, A., Klieme, E., & Köller, O. (2016). *PISA 2015 - Eine Studie zwischen Kontinuität und Innovation*. Münster: Waxmann.

Robinson-Cimpian, J. P., Lubienski, S. T., Ganley, C. M., & Copur-Gencturk, Y. (2014). Teachers' perceptions of students' mathematics proficiency may exacerbate early gender gaps in achievement. *Developmental Psychology*, 50, 1262–1281. https://doi.org/10.1037/a0035073

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

Schöne, C. (2014). Fähigkeitsselbstkonzept. In M. A. Wirtz (Hrsg.), *Dorsch – Lexikon der Psychologi*e (18. Aufl., S. 521). Bern: Verlag Hogrefe Verlag.

Schmitz, B., & Skinner, E. (1993). Perceived control, effort, and academic performance: Interindividual, intraindividual, and multivariate time-series analyses. *Journal of Personality and Social Psychology*, 64(6), 1010. https://doi.org/10.1037/0022-3514.64.6.1010

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

Seidel, T. (2006). The role of student characteristics in studying micro teaching-learning environments. *Learning Environments Research*, 9, 253–271. https://doi.org/10.1007/s10984-006-9012-x

Seidel, T., Prenzel, M., Duit, R., & Lehrke, M. (2003). *Technischer Bericht zur Videostudie "Lehr-Lern-Prozesse im Physikunterricht"* (1., Aufl.). Kiel: IPN Leibniz-Institut für die Pädagogik der Naturwissenschaften an der Universität Kiel.

Solga, H., & Pfahl, L. (2009). Doing Gender im technisch-naturwissenschaftlichen Bereich. In J. Milberg (Hrsg.), *Förderung des Nachwuchses in Technik und Naturwissenschaft* (S. 155–218). Heidelberg: Springer.

Spinath, B. (2005). Akkuratheit der Einschätzung von Schülermerkmalen durch Lehrer und das Konstrukt der diagnostischen Kompetenz. *Zeitschrift für Pädagogische Psychologie*, 19, 85–95. https://doi.org/10.1024/1010-0652.19.12.85

Stang-Rabrig, J., & Urhahne, D. (2016). Wie gut schätzen Lehrkräfte Leistung, Konzentration, Arbeits- und Sozialverhalten ihrer Schülerinnen und Schüler ein? Ein Beitrag zur diagnostischen Kompetenz von Lehrkräften. *Psychologie in Erziehung Und Unterricht*, 63, 204. https://doi.org/10.2378/peu2016.art18d

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. https://doi.org/10.1037/a0027627

Südkamp, A., Möller, J., & Pohlmann, B. (2008). Der Simulierte Klassenraum: Eine experimentelle Untersuchung zur diagnostischen Kompetenz. *Zeitschrift für Pädagogische Psychologie*, 22, 261–276. https://doi.org/10.1024/1010-0652.22.34.261

Terhart, E. (2009). Erste Phase: Lehrerbildung an der Universität. In O. Zlatkin-Troitschanskaia (Hrsg.), *Lehrprofessionalität. Bedingungen, Genese, Wirkungen und ihre Messung* (S. 425–437). Weinheim: Beltz.

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

Tiedemann, J. (2000). Gender related beliefs of teachers in elementary school mathematics.

Educational Studies in Mathematics, 41, 191–207. https://doi.org/10.1023/A:1003953801526

Urhahne, D., Timm, O., Zhu, M., & Tang, M. (2013). Sind unterschätzte Schüler weniger leistungsmotiviert als überschätzte Schüler? *Zeitschrift für Entwicklungspsychologie und Pädagogische Psychologie*, 45, 34–43. https://doi.org/10.1026/0049-8637/a000079

Urhahne, D., Zhou, J., Stobbe, M., Chao, S.-H., Zhu, M., & Shi, J. (2010). Motivationale und affektive Merkmale unterschätzter Schüler. Ein Beitrag zur diagnostischen Kompetenz von Lehrkräften. *Zeitschrift für Pädagogische Psychologie*, 24, 275–288. https://doi.org/10.1024/1010-0652/a000021

van Ophuysen, S. (2006). Vergleich diagnostischer Entscheidungen von Novizen und Experten am Beispiel der Schullaufbahnempfehlung. *Zeitschrift für Entwicklungspsychologie und Pädagogische Psychologie*, 38(4), 154–161. doi:10.1026/0049-8637.38.4.154

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

Wigfield, A. & Eccles, J. S. (2002). The development of competence beliefs, expectancies for success, and achievement values from childhood through adolescence. In A. Wigfield & J. S. Eccles (Hrsg.), *Development of achievement motivation* (S. 173–195). San Diego, CA: Academic Press.

Willems, K. (2007) *Schulische Fachkulturen und Geschlecht: Physik und Deutsch- natürliche Gegenpole?* Bielefeld: Transkript Verlag.

Woodcock, S., & Vialle, W. (2011). Are we exacerbating students' learning disabilities? An investigation of preservice teachers' attributions of the educational outcomes of students with learning disabilities. *Annals of Dyslexia*, 61(2), 223–241. https://doi.org/10.1007/s11881-011-0058-9

Zhu, M., Urhahne, D., & Rubie-Davies, C. M. (2018). The longitudinal effects of teacher judgement and different teacher treatment on students' academic outcomes. *Educational Psychology*, 38(5), 648-668. https://doi.org/10.1080/01443410.2017.1412399

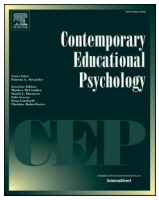
MANUSCRIPT 2

Contemporary Educational Psychology 65 (2021) 101965

Contents lists available at ScienceDirect

Contemporary Educational Psychology

journal homepage: www.elsevier.com/locate/cedpsych



Teacher judgments at zero-acquaintance: A social accuracy analysis

Caroline V. Bhowmik [a,*], Steffen Nestler [b], Friedrich-Wilhelm Schrader [a], Anna-Katharina Praetorius [c], Jeremy C. Biesanz [d], Mitja D. Back [b]

[a] Department of Psychology, University of Koblenz-Landau, Fortstraße 7, 76829 Landau, Germany
[b] University of Münster, Germany
[c] University of Zurich, Switzerland
[d] University of British Colombia, Canada

ARTICLE INFO

Keywords:
Teacher judgment accuracy
Social accuracy model
Academic self-concept
Teacher expectations
Teacher first impressions

ABSTRACT

How accurate are teachers' first impressions and what moderates the degree of first impression accuracy? In previous teacher judgment accuracy research, teachers judged students who were well-acquainted to them, focusing on single traits. Here, we follow the zero-acquaintance paradigm and apply the Social Accuracy Model (SAM; Biesanz, 2010) to examine teachers' first impressions regarding students' personality profiles. Three groups of perceivers (student teachers, experienced teachers and psychology students; $N = 285$) rated students' ($N = 10$) academic self-concept, intrinsic motivation and intelligence based on brief videos. SAM analyses revealed that teachers were accurate regarding the average students' profile of characteristics (normative accuracy), but were not successful at detecting students' unique personality profiles (distinctive accuracy). Moreover, likeable students and those evaluated as more physically attractive were perceived with higher normative accuracy. Personality similarity and teaching experience were unrelated to accuracy. Implications for teacher judgment accuracy research and educational practice are discussed.

1. Introduction

Teacher judgments about their students, which represent a broad spectrum, ranging from reflected appraisals to quick and intuitive impressions, are both widespread and consequential (Artelt, 2016). They fulfill many functions such as assessment and evaluation or fostering students' learning and development, and have the potential to influence students' academic development and outcomes in many respects (Kriegbaum, Steinmayr, & Spinath, 2019; Parks & Kennedy, 2007). Accuracy of teacher judgments is important because validity and fairness of assessment as well as quality of teaching and learning are involved. Currently, accuracy of teacher judgments is mostly dealt with according to two lines of research: First, accurate teacher judgments with respect to students' achievement and their cognitive and motivational characteristics are considered as an essential prerequisite for adaptive teaching (e.g., the selection of learning material and levels of difficulty) and optimizing students' achievement and development (Corno, 2008). Second, judgment accuracy and its relation to bias, i.e. inaccurate perceptions and expectations of teachers, which have a potential to act as *self-fulfilling prophecies* (also referred to as *interpersonal expectancy effects*; Brophy, 1983; Harris & Garris, 2008). As Jussim and Harber (2005) make clear, self-fulfilling prophecies with their sometimes favorable, but often harmful effects can only arise when teacher perceptions are inaccurate. Clarifying origins and determinants of teachers' judgment accuracy, as well as processes underlying accuracy are therefore important research goals. In this study, we focus on teacher judgments that are based on first impressions.

In research on teacher judgment accuracy, teachers are typically asked to judge students from their own classrooms, that is, students whom they are familiar with. One issue that is largely unexplored is the accuracy of teacher judgments about students whom they are unacquainted with and who are only present for a short period of time. Research in personality and social psychology regarding impressions based on minimal information and no prior acquaintance (i.e., first impression research) has demonstrated that judgment accuracy can be remarkably high in such cases. There is much evidence that first impressions have a substantial influence on the way subsequent information is recorded and processed (Anderson & Barrios, 1961; Asch, 1946;

* Corresponding author.
E-mail addresses: wahle@uni-landau.de (C.V. Bhowmik), steffen.nestler@uni-muenster.de (S. Nestler), schrader@uni-landau.de (F.-W. Schrader), anna.praetorius@ife.uzh.ch (A.-K. Praetorius), jbiesanz@psych.ubc.ca (J.C. Biesanz), mitja.back@uni-muenster.de (M.D. Back).

https://doi.org/10.1016/j.cedpsych.2021.101965

Available online 10 March 2021

Darley & Fazio, 1980; Nickerson, 1998; Wason, 1960). Consequently, the accuracy of teacher first impressions is important because inaccurate impressions can lead to biased expectations, erroneous judgments, and, as a consequence, to unfavorable student outcomes.

The present study investigates the accuracy of teachers' rapid judgments regarding students' academic self-concept, motivation and intelligence, which constitute highly relevant factors for students' learning processes and academic outcomes (Furnham & Monsen, 2009; Kriegbaum, Jansen, & Spinath, 2015; Machts, Kaiser, Schmidt, & Möller, 2016; Marsh, 1992; Schunk & Zimmerman, 2012; Skinner & Belmont, 1993). To pursue this aim, we employ a thin-slices and zero-acquaintance approach (Ambady, Hallahan, & Rosenthal, 1995), a methodology broadly applied in personality and social psychology research. In this approach, participants are presented with brief videos or pictures (*thin-slices of behavior*, i.e. short periods of behavior or behavioral episodes lasting from some seconds to a few minutes; Ambady & Rosenthal, 1992; Back & Nestler, 2016) of targets they do not know (*zero-acquaintance*), and are then asked to judge the values of those targets on different traits. By focusing on judgments based on minimal information and no prior acquaintance (e.g., brief video clips of targets), this approach allows to isolate the accuracy of judgment processes per se (i.e., independent of perceiver-target interactions or previous knowledge about targets).

As previous research has revealed large differences in accuracy across teachers (Lorenz, 2011; Lorenz & Artelt, 2009), the present work also explores variables that can moderate judgment accuracy of judgments at zero-acquaintance. Informed by personality and social psychology accuracy research, we examine whether the attractiveness of the students, liking, and personality similarity also play a role in spontaneous teacher judgments, particularly because teachers' first impressions can influence student assessment, which is part of teachers' professional role. Although teachers are expected to make professional judgments, we assume that their judgments can be influenced by factors that play a role in everyday impression formation, particularly when dealing with first impressions of students.

1.1. The accuracy of teacher judgments

In research on teacher judgment accuracy, teachers are usually asked to indicate their judgments regarding the students in their classrooms with respect to certain characteristics or traits. Those judgments are then compared with students' actual characteristics measured by tests or self-reports by calculating accuracy scores (e.g., classwise correlations) between teacher judgments and students' actual characteristics.

Research on this topic is most commonly characterized by a *variable-centered* or *trait-based approach.* In this approach, researchers are interested in discriminating among several students on a particular trait of interest (i.e., inter-target comparison). Most previous studies have focused on students' academic achievement as a criterion (Südkamp, Kaiser, & Möller, 2012). Meta-analyses show that on average teachers have a high ability to accurately perceive students' academic achievement. Südkamp et al. (2012), for example, found an average correlation of $r = .63$ across 75 studies, which can be interpreted as a strong effect (Funder & Ozer, 2019; Gignac & Szodorai, 2016). Judgment accuracy levels, however, vary considerably between studies ($-.03 \leq r \leq .83$[1]; e.g., Südkamp et al., 2012) and individual teachers (e.g., $.11 \leq r \leq .88$; Lorenz & Artelt, 2009).

A smaller amount of studies has dealt with cognitive abilities (e.g., intelligence), academic self-concept and motivation. These characteristics are closely related to achievement but are also associated with basic personality traits that have been studied in personality research. For students' cognitive ability, a recent meta-analysis by Machts et al. (2016) revealed an average accuracy correlation of $r = .50$ based on 33 studies. For the academic self-concept, mean correlations between students' self-reports and teacher judgments are usually smaller and moderate to large in size ($.29 \leq r \leq .55$; Praetorius et al., 2015; Praetorius, Karst, Dickhäuser, & Lipowsky, 2011; Spinath, 2005).

In the case of students' intrinsic motivation, correlations between teacher judgments and students' actual characteristics are small to moderate ($.10 \leq r \leq .20$; Spinath, 2005; Urhahne, Chao, Florineth, Luttenberger, & Paechter, 2011). As it is the case for academic performance, accuracy levels for teacher judgments of students' academic self-concept and motivation vary considerably between individual teachers (academic self-concept: $-.45 \leq r \leq .89$; Praetorius et al., 2015; Praetorius, Karst, Dickhäuser, & Lipowsky, 2011; Spinath, 2005; motivation: $-.46 \leq r \leq .67$; Spinath, 2005).

Contrary to the trait-based approach, *person-centered* or *profile-based approaches* compare the agreement between judgment and criterion for several different traits *within* one target (i.e., intra-target comparison). It is important to distinguish between both approaches because they refer to different and potentially independent psychological phenomena (Back & Nestler, 2016; Hall et al., 2017). The fact that teachers usually need to perceive and evaluate several student characteristics simultaneously supports the potential of applying a profile-based approach when investigating teacher judgment accuracy (Südkamp, Praetorius, & Spinath, 2018).

For teacher judgments, two components of accuracy are especially important, one of them relating to single students, and the other one to groups of students. The first kind of judgment enables teachers to tailor their teaching to individual students' aptitudes. The second one supports teachers to adapt their instruction to the learning requirements of the whole class or age group. In much the same manner, teacher expectations can refer to individual students (as was the case in early research by Brophy & Good, 1974) or to the whole class (becoming a topic of later research by Brophy & Good, 1984).

In light of the above, profile analyses allow disentangling two main components of judgmental accuracy: *normative accuracy* (i.e., knowledge about how the average person is like) and *distinctive profile accuracy* (i.e., knowledge about an individual's unique personality profile) (Bernieri, Zuckerman, Koestner, & Rosenthal, 1994; Biesanz, 2010, 2019; Borkenau & Leising, 2016; Cronbach, 1955; Hall et al., 2017). In the context of teacher judgments, normative accuracy reflects teachers' more or less accurate knowledge of the characteristics of a student group as a whole that is important for classroom-based decisions and actions. Distinctive accuracy, on the other hand, is important for the treatment of individual students within a classroom. Hence, by offering insights regarding both key aspects of accuracy, the profile-based approach bears a large potential for research on teacher judgments.

1.2. The accuracy of personality judgments at zero-acquaintance based on thin slices of behavior

Whereas teacher judgment research is usually conducted in classroom settings with existing acquaintance and prior interaction between perceivers (teachers) and targets (students), judgment accuracy research in personality and social psychology is often based on a zero-acquaintance or thin-slices approach (Ambady & Rosenthal, 1992; Ambady et al., 1995; Connelly & Ones, 2010), which also provides a suitable framework to examine the accuracy of first impressions that teachers form when they meet a student for the first time. First impressions often reflect stereotypes, which are defined as persons' sets of beliefs about the personal attributes of a certain social group and characterize expectations about people or groups of people in general (see also Ashmore & Del Boca, 1981). Normative accuracy (also: *stereotype accuracy* according to Cronbach, 1955) concerns the correspondence between a person's knowledge about the average target and the general pattern of actual target attributes in a group or population whereas distinctive accuracy characterizes knowledge about how an

[1] Values were transformed back to correlations from Fisher's z-scores as reported by Südkamp, Kaiser, & Möller (2012)

individual target deviates from this general pattern.

Previous research in personality and social psychology using the *Big Five* traits (conscientiousness, agreeableness, neuroticism, openness to experience, extraversion) showed that individuals can form impressions of other individuals' personality rapidly and with a considerable degree of accuracy (Ambady & Rosenthal, 1992). In trait-based analyses, accuracy levels usually range between $r = .10$ and $r = .40$ across traits and contexts (Back & Nestler, 2016; Connelly & Ones, 2010; Hall, Andrzejewski, Murphy, Mast, & Feinstein, 2008; Kenny & West, 2010). Judgments of other characteristics (e.g., intelligence) show similar amounts of accuracy (Borkenau, Leising, & Fritz, 2014; Murphy, 2007; Murphy, Hall, & Colvin, 2003; Reynolds & Gifford, 2001). Profile-based approaches have shown comparable ranges of accuracy, whereby normative accuracy levels usually substantially outperform distinctive accuracy outcomes (Back & Nestler, 2016; Borkenau & Leising, 2016).

A first study applying a thin-slice or zero-acquaintance approach in the teacher judgment context investigated teachers' accuracy in judging the academic self-concept of unacquainted students based on 30-second videos (Praetorius et al., 2015). Following a variable-centered or trait-based approach, classwise correlations between teacher judgments and students' actual characteristics were calculated. The average accuracy in the zero-acquaintance condition was $r = .35$ (with a range of $.31 \leq r \leq .39$) across four groups of judges. This result shows that teachers were able to judge the rank order of unacquainted students to some degree. Moderators of zero-acquaintance accuracy were not included.

1.3. Moderators of judgment accuracy

Whereas our knowledge of potential moderators for teacher judgment accuracy is to date very limited and confined to the variable-centered approach and classroom research with familiar students (Südkamp et al., 2012), the investigation of moderators of judgment accuracy is an integral part of personality and social psychology research. In this area, several characteristics of the judge have been reported to promote judgment accuracy, among which are intelligence, psychological adjustment and interpersonal experience (Davis & Kraus, 1997; Letzring, 2008, 2010). But empirical evidence across studies is mixed and inconsistent (Back & Nestler, 2016). In the context of education, the counterpart to interpersonal experience is teaching experience. However, for teaching experience, which is often considered as an indicator of teaching expertise (Berliner, 2001), a relation with accuracy could so far not be confirmed (Herppich, Wittwer, Nückles, & Renkl, 2013; Ready & Wright, 2011). Results could be different when applying a profile-based approach, because obtaining accurate judgments with respect to several traits simultaneously is a more demanding task that could therefore benefit from teachers' experience in such complex situations.

Characteristics of the target as well as variables that refer to the individual perceiver-target dyad (i.e., *dyadic moderators*) may also moderate judgmental accuracy. Given that certain judges perceive certain targets more accurately than other targets, it matters which judge gets confronted with which target. Personality and social psychology research showed that liking (Human & Biesanz, 2011b; Leising, Ostrovski, & Zimmermann, 2013), gender and ethnic similarity (Letzring, 2010) as well as similarity with respect to personality and attitudes promote accuracy (Byrne, 1997; Montoya, Horton, & Kirchner, 2008; Montoya & Horton, 2013). Furthermore, physical attractiveness conceptualized as a dyadic variable when each target's attractiveness is evaluated by the judge in each judge-target dyad, has been shown to promote first impression accuracy (Lorenzo, Biesanz, & Human, 2010). Contrary to attractiveness as a target attribute that can be objectively measured, such as symmetry of face, or rated with high inter-rater agreement, attractiveness as a dyadic variable is a completely personal and intuitive view of a target, which may be different for different perceivers.

The moderating effect of judge-target similarity, liking and perceived physical attractiveness can be explained in such way that perceivers have deeper knowledge about people who are similar to them or people they like, are more interested in these individuals, and pay closer attention to them (Human, Biesanz, Parisotto, & Dunn, 2012). In the context of first impressions, liking typically refers to a very spontaneous form of interpersonal appeal and is therefore often used interchangeably with the term interpersonal attraction (Back, Schmukle, & Egloff, 2011). Also similarity with respect to personality and attitudes, for example, can lead to an increased attraction towards the target (Byrne, 1997; Montoya et al., 2008; Montoya & Horton, 2013). As a consequence, perceivers are expected to be able to detect more cues and process them in more detail resulting in higher judgment accuracy (Biesanz et al., 2011; Funder, 1995; Lorenzo et al., 2010).

The moderating role of different variables can depend on the amount and quality of information available to the perceivers (also referred to as *cues*). For instance, being similar to the target will be of an advantage, when the target elicits observable cues, which are then differentially perceived by the perceivers according to their own individual personality characteristics (Back et al., 2011). This is consistent with the idea of the realistic accuracy model (Funder, 1995), in which the importance of relevance and availability of behavioral and physical cues for an accurate judgment is described.

Whereas research in personality and social psychology emphasizes the favorable effects of factors such as personality similarity, liking and physical attractiveness on judgment accuracy, from an educational point of view, these characteristics represent sources of bias, which can impair teachers' professional judgment. Therefore, exploring their actual role as moderators of teacher judgment seems to be important, particularly because previous research has shown that teachers are equally susceptible to judgment biases as are people in regular judgment contexts (Parks & Kennedy, 2007). Students' physical attractiveness, for example, has been shown to be associated with more favorable ratings by the teachers on a number of dimensions, such as intelligence, academic potential, grades and social skills (Clifford & Walster, 1973; Parks & Kennedy, 2007; Ritts, Patterson, & Tubbs, 1992). Thus, teachers' impressions of students' physical attractiveness can be seen as a relevant source of teacher expectations with respect to both academic performance and social/personality attributes (Clifford & Walster, 1973; Dusek & Joseph, 1983). It is therefore important to examine if such factors also affect the accuracy of teachers' first impressions.

1.4. Analyzing teacher judgment accuracy: The Social Accuracy Model (SAM)

SAM is a statistical model that is widely applied in personality and social psychology research with a focus on investigating personality judgments and influencing factors for accuracy (Biesanz et al., 2011; Human et al., 2013, 2014; Human & Biesanz, 2011a, 2011b; Letzring, 2015; Letzring & Human, 2014).

To investigate profile accuracy and its moderators, SAM (see Biesanz, 2010, 2019 for a detailed description of the model and application examples) is a promising approach as it allows to simultaneously examine such effects in a straightforward mathematical model. SAM integrates Cronbach's (1955) componential approach with Kenny's Social Relations Model (SRM; 1994). The model allows examining the degree and moderators of profile accuracy and thereby disentangling the two main components: normative accuracy and distinctive accuracy. In the school context, these two accuracy components describe teachers stereotypic expectations (i.e., normative accuracy) and their perceptions of individual students' uniqueness (i.e. distinctive accuracy).

Besides analyzing the degree of both types of judgment accuracy, SAM can be used to examine whether judge variables (e.g., teaching experience) as well as judge-target relationship variables (e.g., similarity) affect the two accuracies and thereby moderate the association between actual and perceived profiles. Given that a larger number of different judgment variables are examined simultaneously, SAM offers

substantial statistical power even in cases where only few target participants are available.

1.5. The present research

The present research aims at analyzing perceivers' judgment accuracy regarding students' trait profile based on minimal information and investigates moderators affecting judgment accuracy. We focus on judgments of students' intelligence, academic self-concept and intrinsic motivation as these characteristics have shown to play a crucial role for students' academic achievement and success (Kriegbaum et al., 2015; Lipnevich & Roberts, 2012; Machts et al., 2016; Möller, Pohlmann, Köller, & Marsh, 2016). We use an instructional context in which students work on a physics experiment independently as this setup provides an opportunity to observe a greater amount of highly visible cues than in seatwork or group discussions. Based on the same experimental situation for each student, we moreover ensured a constant setting in which the variability was kept as low as possible. Applying a zero-acquaintance and thin-slice of behavior approach, we try to capture some important features of the formation of first impressions that translate to real situations when teachers meet their students for the first time. To explore the degree of teacher judgment accuracy, a profile-based approach using the SAM is chosen that allows us to examine teacher judgment accuracy regarding students' profile, i.e. several student characteristics simultaneously and to additionally analyze moderating effects of liking, perceived attractiveness, and personality similarity.

We address the following research questions:

1 How accurate are perceivers in detecting students' distinctive and students' average trait profile based on minimal information (i.e., brief videos) and no prior acquaintance? Does accuracy vary across different groups of perceivers (experienced teachers, teacher students, psychology students)?
2 What roles do liking, perceived physical attractiveness, and perceiver-student personality similarity play for judgment accuracy?

2. Method

2.1. Procedure

The research design of the present study is based on the large existing body of research employing a thin-slice of behavior approach in personality and social psychology using brief videos of targets as stimulus material. Perceivers watched 30-second videos of ten targets, showing one target in each brief video while working on a physics experiment task during a school visit at the university laboratory. Right after watching a target video, perceivers rated the self-concept, intelligence and motivation of each target. To control for sequence effects, the order of the videos was randomized. Thereafter, perceivers were asked to indicate liking and physical attractiveness of each target. Subsequent to the completion of all ratings, perceivers filled in a personality inventory. Criterion variables and personality characteristics of the students were assessed prior to their visit.

2.2. Target subjects

Students ($n = 10$, 50% female, $M_{age} = 15.6$, $SD_{age} = 0.84$) in tenth grade served as targets in this study. They were part of a total sample of $N = 244$ students from ten German secondary schools who were visiting the university laboratory for working on physics experiments. Students in the target sample came from four of the ten schools and were selected in such a way that differences between targets in the criterion variables should be recognizable; they were thus chosen so that they displayed a high variance of students' intelligence, academic self-concept in physics, general academic self-concept, and intrinsic motivation (see Table 1 for SD and range of all criterion variables). Moreover, quality and comparability of the video snippets were important (i.e., mostly frontal gazing student, no interaction with other students and/or the teacher and a comparable resolution and light). All students were unacquainted to the perceivers.

2.3. Perceiver subjects

Overall, $N = 285$ perceivers participated in this study. The first group of perceivers consisted of $N = 104$ undergraduate students enrolled in the teacher study program at two German universities (64.3% female, $M_{age} = 22.04$ years, $SD_{age} = 4.41$). The second group of perceivers consisted of $N = 99$ experienced teachers (59.78% female, $M_{age} = 36.01$ years, $SD_{age} = 9.22$; $M = 8.6$ years of professional experience, $SD = 8.7$; min. = 1; max. = 45). Of this group, the majority of teachers (83.16%) were secondary school teachers. The third group consisted of $N = 82$ undergraduate psychology students at a German university (77.5% female, $M_{age} = 22.39$ years, $SD_{age} = 5.1$). While experienced teachers received a 25-euro voucher for their participation, both psychology and teacher students participated in exchange for course credit.

2.4. Criterion measures

Students' intrinsic motivation for physics was measured using the PISA 2006 (Frey et al., 2009) scale (e.g., "I enjoy attending physics lessons in school") comprising three items. Answering categories for intrinsic motivation ranged from 1 ("never") to 4 ("almost always"). To obtain a more complete estimate of students' academic self-concept, we included both the general and subject-specific self-concept in the study.

Table 1
Means (M), standard deviations (SD), minimum (Min), maximum (Max), internal consistencies (α), and intercorrelations of targets' characteristics.

Criterion variables	*M*	*SD*	*Min*	*Max*	*α*	1	2	3
1. Self-concept								
Raw scores	4.25	0.83	2.71	5.43				
Pomp scores	0.65	0.10	0.44	0.80	.80	–		
2. Self-concept physics								
Raw scores	2.70	0.63	1.75	4.00				
Pomp scores	0.57	0.20	0.25	1.00	.85	.69*	–	
3. Intrinsic motivation physics								
Raw scores	2.43	0.59	1.67	3.33				
Pomp scores	0.48	0.06	0.40	0.53	.89	.28	.56	–
4. Intelligence								
Raw scores	113.05	16.13	86.50	128.50				
Pomp scores	0.64	0.26	0.20	0.90		.65*	.74*	.27

Note. * $p < .05$. $N_{Targets} = 10$. Raw intelligence scores were transformed into IQ scores based on the test manual of the IST-Screening. Internal consistencies (α) refer to the complete student sample ($N = 244$) and are based on pomp scores (see also 2.5. for a description of the applied data transformation method). For intelligence, test validation studies by the authors of the IST-Screening demonstrate reliabilities between .72 and .90 (Cronbach's alpha, split half, test–retest).

Students' academic self-concept in physics served as the subject-specific self-concept and was measured using a 4-item scale (e.g., "I am gifted in physics"; Seidel, Prenzel, Duit, & Lehrke, 2003). Answering categories ranged from 1 ("do not agree") to 4 ("fully agree"). To assess students' general academic self-concept, the DISK-Gitter (Rost, Sparfeldt, & Schilling, 2007) containing eight items was used (e.g., "I receive good grades at school") with categories ranging from 1 ("do not agree") to 6 ("fully agree"). Students' intelligence was assessed using the IST-Screening (Liepmann, Beauducel, Brocke, & Nettelstroh, 2012). This test consists of three subtests: word analogies, numerical series and matrices.

2.5. Perceiver judgments

Perceivers rated students' intelligence and academic self-concept in physics based on a 1-item scale ("In your opinion, how intelligent is this student?" and "How do you think this student is estimating his/her skills in physics?"). Before judging students' academic self-concept in physics, however, perceivers were presented with the four items of the student questionnaire to get familiarized with the items the students were presented with (i.e., *informed judgment*; Südkamp et al., 2012). Judgment categories ranged from 1 ("very low") to 4 ("very high") for the academic self-concept in physics and 1 ("not intelligent") to 6 ("very intelligent") for students' intelligence. Students' intrinsic motivation was rated based on the identical three items that were applied in assessing students' self-reports (e.g., "The student enjoys attending physics lessons in school") with the 4 - point scale ranging from 1 ("fully disagree") to 4 ("fully agree"). The same applies to the ratings of students' general academic self-concept, where the judgments were based on the original seven items (e.g., "The student receives good grades"), with answering categories ranging from 1 ("do not agree") to 6 ("fully agree'). As one item of the DISK-Gitter ("I belong to the good ones in school") was decreasing the reliability of the scale considerably, this item was excluded from the analysis to result in a total of 12 instead of 13 items. Thus, perceivers rated students' characteristics on a total of 12 items. Consequently, the item was removed from students' criterion data as well. In order to standardize the values obtained from varying rating and criterion measures, we used a transformation equivalent to *Percent of Maximum Possible (POMP) Scores* (Cohen, Cohen, Aiken, & West, 1999; Fischer & Milfont, 2010). Transformed scores were obtained by taking the raw score of each item minus the minimum score and then dividing it by the possible scoring range. For example, for students' intrinsic motivation with a range of values from 1 to 4, for each of the three items the minimum score (1) was subtracted from students' self-report on that item before dividing it by 3 (possible scoring range). The same was done for the perceiver ratings. The transformed values were then ranging from 0 to 1 for all student and perceiver variables. For students' intelligence, POMP equivalent scores were calculated based on the minimum and maximum intelligence scores of the larger student sample ($N = 244$), from which the ten targets were selected.

2.6. Moderator variables

Teaching experience was operationalized via the three sample groups. To investigate differences in accuracy between the three groups, we computed one dummy variable that coded teachers versus teacher students and a second dummy variable that coded psychology students versus teacher students. This dummy coding enables the comparison between teachers (i.e., large amount of teaching experience), as well as psychology students (i.e., no teaching experience) with the teacher students as our baseline group (i.e., small to medium amount of teaching experience) with respect to judgment accuracy. One item was used to assess perceivers' liking for a target ("How much do you like this particular student?"). Another item measured perceivers' perception of each target's physical attractiveness ("In your opinion, how attractive is this student?"). Both items had to be answered on a scale ranging from 1 ("not at all") to 6 ("very much"). Both liking and attractiveness are defined as dyadic variables resulting in scores that are specific for each perceiver-target dyad. When assessing students' physical attractiveness, we were interested in perceivers' personal evaluation regarding how good looking they perceive individual students.

To calculate personality similarity scores, we used a perceiver's and a student's answers to a 10-item version of the Big Five Inventory (BFI-10; Rammstedt & John, 2007). Each of the ten items had to be answered on a 5-point Likert Scale ranging from 1 (completely disagree) to 5 (completely agree). With two items each, the BFI-10 measures the five personality dimensions, i.e., extraversion, neuroticism, agreeableness, conscientiousness, and openness resulting in a 10-variable profile for each perceiver and each target. We computed a (profile) correlation to measure the similarity between the two ten-item profiles of a perceiver and a target. This correlation coefficient was calculated for each perceiver-target dyad resulting in 2850 unique similarity scores that were used as a variable at the dyad level.

3. Analyses

3.1. Trait-based accuracy

While the main focus of the present paper lies on profile-based accuracy analyses, we additionally provide trait-based accuracy analyses for descriptive purposes. To obtain trait-based accuracy scores, we calculated Pearson correlations between the perceiver ratings and the target data separate for each criterion and perceiver in the study. Single perceiver correlations were then Fisher's z transformed, averaged, and then back-transformed to correlations to obtain average judgment accuracy values separate for each criterion variable (see Back & Nestler, 2016).

3.2. Social accuracy analyses

SAM is a multilevel model with the dyad as the basic unit of analysis. Each dyad (each specific perceiver-target combination) is described by corresponding profiles of judged and actual traits (items). Normative and distinctive accuracy are profile-based accuracies based on a regression of a perceiver's judgments (dependent variable) on actual traits (independent variable) calculated over traits or items for each dyad. The slopes of this regression indicate distinctive accuracy (with individual target values as predictor) and normative accuracy (with mean target values as predictor), and the intercept represents a level parameter. Pooling and aggregating these dyad-specific regressions over dyads is statistically problematic as it does not account for the dependencies in the data. For instance, one source of a dependency is that judgments of an individual perceiver are more similar to each other than judgments across all perceivers. As a form of a multilevel model, SAM takes these dependencies into account. The fixed effects part of the model describes the mean relation between judgments and criteria, averaged over all dyads in the sample. The random effects part represents the variability of the dyadic (teacher × students) regressions and can be split into variance accounted by perceivers (perceptive accuracy, targets (expressive accuracy), and perceiver-target interactions (dyadic variability). In order to analyze moderator effects, moderator variables are added as predictors of the random regression coefficients.

On a mathematical level, SAM is a crossed-random effects multilevel model in which each observation at Level 1 (i.e., a single judgment) is nested within perceivers (as targets are judged by the same set of perceivers) and targets (as perceivers judges the same set of targets). Therefore, compared to standard multilevel models, this model is not strictly hierarchical. In the present case, SAM accounts for the fact that the total of 2850 dyadic judgments at Level 1 are not independent because they are nested in perceivers ($N = 285$) *and* in student targets ($N = 10$). Perceiver ratings serve as the dependent variable and the judgment criterion (i.e., target data) as the independent variable.The

following model was fitted to the data:

$$Y_{ijk} = \beta_{0ij} + \beta_{1ij}TTrait_{jk} + \beta_{2ij}Mean_k + \varepsilon_{ijk} \quad \begin{aligned} \beta_{0ij} &= \beta_{00} + u_{0i} + u_{0j} + u_{0(ij)} \\ \beta_{1ij} &= \beta_{10} + u_{1i} + u_{1j} + u_{1(ij)} \\ \beta_{2ij} &= \beta_{20} + u_{2i} + u_{2j} + u_{2(ij)} \end{aligned} \quad (1)$$

Here, Y_{ijk} corresponds to perceiver i's rating of a target j on item k and $TTrait_{jk}$ represents target j's self-report on item k. The predictor $Mean_k$ is the average self-report of the targets on item k. To obtain a more valid estimate, $Mean_k$ was computed as the average self-report of the complete student sample ($N = 244$), from which the ten targets were selected (see Biesanz, 2010, 2019). β_{0ij} is an intercept term describing a level component, and β_{1ij} denotes the distinctive profile accuracy of perceiver i concerning target j. Lastly, β_{2ij} describes the estimated level of normative agreement for perceiver i with target j –-the congruence between the perceiver's ratings and the average student's self-reported profile after partialling student j 's self-report. The terms β_{00}, β_{10}, and β_{20} refer to the fixed effects in the model, i.e., the intercept, distinctive accuracy, and normative slope, across both targets and perceivers.

Hence, β_{0ij}, β_{1ij}, and β_{2ij} are scores characterizing each single dyad, whereas β_{00}, β_{10}, and β_{20} refer to mean values calculated across all dyads.

The u's correspond to the random effects in the model. The variance of these terms describes the degree of variability around the average levels of distinctive and normative accuracy (reported as *SD*'s), respectively, and the sum of the average fixed effect and a person's unique random effect represents the person-specific effect. Whereas perceptive accuracy refers to the individual differences among the perceivers in their judgment accuracy, expressive accuracy refers to the individual differences among targets in their judgability. Finally, dyadic variability refers to the differences in the agreement between perceiver ratings and target characteristics among the perceiver-target dyads across the 12 items. Thus, $\beta_{10} + u_{1i}$ describes perceiver i's unique distinctive accuracy slope averaged across the ten targets in this study (perceptive accuracy), while $\beta_{10} + u_{1j}$ describes target j's unique distinctive accuracy slope averaged across the 285 perceivers (expressive accuracy). Similarly, $\beta_{20} + u_{1i}$ represents perceiver i's unique normative accuracy slope averaged across the ten targets (perceptive accuracy), while $\beta_{20} + u_{2j}$ indicates target's j's normative accuracy slope averaged across the 285 perceivers (expressive accuracy). Lastly, $u_{1(ij)}$ and $u_{2(ij)}$ represent the dyadic component of the model (dyadic variability), that is, the deviation of perceiver i's accuracy in perceiving target j after removing perceiver and target main effects (u_{1i} and u_{1j}). Thus, dyadic variability refers to interactions between perceivers and targets, accounting for the fact that accuracy can be particularly high or low for different perceiver-target-combinations.

To test whether distinctive and normative judgment accuracy are related to the perceiver (i.e., teaching experience) or variables of the target-perceiver relationship (i.e., similarity, liking and attractiveness of each target as indicated by the perceivers), these moderator variables were added as predictors of the random regression coefficients (and the intercept to control for main effects) to the model.

The respective equations were:

$$Y_{ijk} = \beta_{0ij} + \beta_{1ij}TTrait_{jk} + \beta_{2ij}Mean_k + \varepsilon_{ijk} \quad \begin{aligned} \beta_{0ij} &= \beta_{00} + \beta_{01}Mod_{ij} + u_{0i} + u_{0j} + u_{0(ij)} \\ \beta_{1ij} &= \beta_{10} + \beta_{11}Mod_{ij} + u_{1i} + u_{1j} + u_{1(ij)} \\ \beta_{2ij} &= \beta_{20} + \beta_{21}Mod_{ij} + u_{2i} + u_{2j} + u_{2(ij)} \end{aligned} \quad (1.1)$$

The complete mixed model, which is obtained by substituting the β-weights in the first line of Eq. (1.1), is depicted in Eq. (1.2).

$$\begin{aligned} Y_{ijk} = {} & \beta_{00} + \beta_{01}Mod_{ij} + \beta_{10}TTrait_{jk} + \beta_{20}\ Mean_k \\ & + \beta_{11}Mod_{ij} \times TTrait_{jk} + \beta_{21}Mod_{ij} \times Mean_k \\ & + \left(u_{1i} + u_{1j} + u_{1(ij)}\right)TTrait_{jk} + \left(u_{2i} + u_{2j} + u_{2(ij)}\right)Mean_k \\ & + u_{0i}\ + u_{0j} + u_{0(ij)} + \varepsilon_{ijk} \end{aligned} \quad (1.2)$$

The upper two lines refer to the fixed-effects part of the model, which comprises the analyses of the moderator effects, and the lower two lines refer to the random effects. Here Mod_{ij} is examined as a potential moderator for distinctive accuracy (β_{11}) as well as normative accuracy (β_{21}). The model parameter Mod_{ij} concerns the moderators liking, physical attractiveness and personality similarity as perceivers (i) were estimating each targets' (j) likeability and attractiveness and hence, those two moderators are dyadic moderators. As an estimate that is derived from correlating perceiver and target scores on the Big Five, personality similarity is also a dyadic moderator. For teaching experience, the model parameter would correctly be specified as Mod_i, as this moderator only concerns the perceivers (i). Hence, the unstandardized coefficients β_{11}, and β_{21} (reported as b's) represent the interaction between each moderator variable and distinctive accuracy and normative accuracy, respectively. For example, for the case of our moderator variable liking, positive values on these estimates would indicate that targets that were liked by the perceivers were viewed with greater distinctive and normative accuracy.

Targets' self-reports and intelligence test scores, as well as perceivers' indications of liking and target attractiveness, were centered around the grand mean prior to the analysis.

All models were estimated with the statistical software *R*, using R's *lme4* package (Bates, Mächler, Bolker, & Walker, 2015; R Development Team, 2020). The data as well as the R codes can be downloaded from https://osf.io/ynemd/.

4. Results

4.1. Descriptive statistics and preliminary analyses

4.1.1. Target self-reports and perceiver ratings

Descriptive statistics for all target self-report variables can be found in Table 1 and for the perceiver ratings in Table 2.

4.1.2. Moderator variables

Descriptive values and intercorrelations of all the three moderator variables are presented in Table 3. Notably, personality similarity showed a strong variability across all perceiver-target dyads. Moreover, perceivers' evaluations of students' physical attractiveness and their evaluation of students' likeability showed a high correlation, but none of these two variables did correlate with personality similarity.

4.1.3. Trait-based accuracy

Trait-based accuracy scores can be found in Table 4. Whereas the highest level of judgment accuracy was achieved for students' academic self-concept in physics, the other three constructs were less accurately perceived, however, all at a comparative level. On a descriptive level, no

Table 2
Means (M), standard deviations (SD), minimum (Min), maximum (Max) and intercorrelations of the perceiver ratings.

Perceiver Ratings	*M*	*SD*	*Min*	*Max*	1	2	3
1. Self-concept							
Raw scores	3.74	0.41	2.13	4.76			
Pomp scores	0.55	0.12	0.08	0.90	–		
2. Self-concept physics							
Raw scores	2.78	0.38	1.00	4.00			
Pomp scores	0.45	0.21	0.00	0.75	.61**	–	
3. Intrinsic motivation physics							
Raw scores	2.39	0.31	1.52	3.60			
Pomp scores	0.46	0.11	0.11	0.87	.39**	.41**	–
4. Intelligence							
Raw scores	4.11	0.45	2.70	5.30			
Pomp scores	0.62	0.19	0.00	1.00	.52**	.56**	.40**

Note. $^{**}p < .01$. $N_{Total} = 2850$ (10 Targets × 285 Perceivers).

Table 3
Means (M), standard deviations (SD), Minimum (Min) and Maximum (Max), and intercorrelations among the judge-target relationship moderator variables.

Moderator variables	*M*	*SD*	*Min*	*Max*	1	2	3
1. Liking	4.15	0.57	2.20	5.90	–		
2. Attractiveness	3.73	0.58	1.10	5.30	.49**	–	
3. Personality similarity	−0.01	0.58	−.99	.99	−.01	.00	–

Note. **$p < .01$. $N_{Total} = 2850$ (10 Targets × 285 Perceivers). The mean value of personality similarity was Fisher-*z* transformed before averaging and transformed back afterwards.

differences were found between the three sample groups for any of the four constructs, i.e., teaching experience was not related to trait-based accuracy outcomes.

4.2. Social accuracy analyses

Results of the SAM for the complete sample and each of the three sample groups without moderators are presented in Table 5.

4.2.1. Teachers' average profile accuracy (fixed effect estimates)

Average normative accuracy was large and significantly different from zero in the complete sample as well as in all three subsamples ($b = 0.69$, $p < .001$ for the complete sample). This finding indicates that on average, perceivers judged the students consistent in accordance with the normative profile (mean profile of the group's actual traits). Furthermore, we found that average distinctive accuracy was low and non-significant across all samples ($b = 0.03$, $p = .43$). This indicates that on average, perceivers were not able to judge the unique profile of the pupils. Finally, the average distinctive accuracy coefficients were similar across three sample groups (experienced teacher, teacher students, psychology undergraduate students).

4.2.2. Variability of profile accuracy (random effects)

Variability in normative and distinctive accuracies can be due to differences between perceivers (i.e., perceptive accuracy), targets (i.e., expressive accuracy), and dyads (i.e., dyadic variability). For normative accuracy, perceiver differences were quite substantial for the total sample as well as across all three sample groups (estimated *SD* for u_{2i} is 0.30 for the complete sample). Thus, our sample includes teachers with high and low normative accuracy scores. We also found differences in normative accuracy between targets. The highest variance term was found for the group of teacher students (estimated *SD* for $u_{1j} = 0.22$) followed by psychology undergraduate students (estimated *SD* for $u_{1j} = 0.19$). This indicates that some students are quite similar to the average profile of students' traits whereas other students differ from the average profile to some degree. Finally, dyadic variability indicating differences between specific teacher-student combinations was low (estimated *SD* for $u_{2ij} = 0.05$).

For distinctive accuracy, variance terms were lowest for the perceivers ($u_{1ij} = 0.02$), followed by the targets ($u_{1ij} = 0.10$) and the dyads ($u_{1ij} = 0.12$). These results suggest that teachers do not vary in their ability to perceive targets' individual profiles accurately (perceptive accuracy). However, a teacher's distinctive accuracy depends to some degree on the specific student he or she is confronted with (dyadic variability), and the student's judgability (expressive accuracy). For all of the three accuracies, no considerable differences in the random effects were identified between the three sample groups.

Table 4
Trait-based accuracies for each variable and sample group.

	Teachers ($n = 99$)				Teacher students ($n = 104$)				Psychology students ($n = 82$)				Complete sample ($N = 285$)			
Trait-based accuracy	*M*	*SD*	*Min*	*Max*	*M*	*SD*	*Min*	*Max*	*M*	*SD*	*Min*	*Max*	*M*	*SD*	*Min*	*Max*
Self-concept	.15***	0.31	−0.55	0.76	.11***	0.33	−0.53	0.78	.13***	0.32	−0.66	0.80	.13***	0.32	−0.66	0.78
Self-concept physics	.49***	0.28	−0.19	0.81	.42***	0.33	−0.48	0.90	.47***	0.32	−0.25	0.89	.46***	0.31	−0.48	0.90
Intrinsic motivation physics	.19***	0.37	−0.61	0.86	.15***	0.30	−0.76	0.76	.17***	0.37	−0.66	0.71	.17***	0.34	−0.76	0.86
Intelligence	.18***	0.14	−0.24	0.47	.17***	0.14	−0.18	0.49	.16***	0.31	−0.61	0.73	.18***	0.14	−0.77	0.85

Note. ***$p < .001$. $N = 285$ correlations, each based on 10 targets. Single-perceiver trait-based accuracy values were obtained by correlating each perceiver's ratings with targets' self-reports for each construct individually. All correlations are Pearson correlations and the values were Fisher-*z* transformed before averaging and transformed back afterwards. One sample t-tests were applied to test if mean correlations differ significantly from zero.

Table 5
Random and fixed effects estimates from the base model with no interactions.

Model Parameters	Teachers (n = 99)		Teacher students (n = 104)		Psychology students (n = 82)		Complete sample (n = 285)	
	b (SE)	*r*	*b (SE)*	*r*	*b (SE)*	*r*	*b (SE)*	*r*
Fixed Effects Estimates								
Distinctive Accuracy	0.02 (0.03)	.02	0.04 (0.03)	.03	0.02 (0.04)	.02	0.03 (0.03)	.02
Normative Accuracy	0.61 (0.05)***	.36	0.72 (0.08)***	.28	0.75 (0.08)***	.33	0.69 (0.05)***	.23
Random Effects (*SD*)								
Perceiver (Perceptive Accuracy)								
Distinctive Accuracy (u_{1i})	0.02		0.03		0.02		0.02	
Normative Accuracy (u_{2i})	0.30		0.25		0.33		0.30	
Target (Expressive Accuracy)								
Distinctive Accuracy (u_{1j})	0.09		0.10		0.12		0.10	
Normative Accuracy (u_{2j})	0.12		0.22		0.19		0.16	
Dyadic Variability								
Distinctive Accuracy (u_{1ij})	0.14		0.12		0.09		0.12	
Normative Accuracy (u_{2ij})	0.01		0.08		0.09		0.05	
Dyadic Impressions (*N*)	985		1040		820		2845	

Note. ***$p < .001$. Sample size (*n*) refers to the number of perceivers. Fixed effect estimates are unstandardized regression coefficients; *r* refers to correlations based on a standardized approximation of effect sizes of the estimates ($r = \sqrt{t^2/(t^2 + dfs)}$).

4.2.3. Moderators of teacher judgment accuracy

Moderator effects refer to differences in judgment accuracy between different perceiver-target dyads. Average effects (see Eqs. (1.1) and (1.2)) for all moderator variables are displayed in Table 6. Effect sizes were calculated using an approximation of *r* (Borenstein, Hedges, Higgins, & Rothstein, 2011).

4.2.3.1. Teaching experience. Normative accuracy was especially high for teacher students in the baseline condition ($b_{20} = 0.72$, $p < .01$), and even significantly higher than for experienced teachers ($b_{21} = -0.13$, $p < .05$), whereas accuracies for teacher students and psychology students did not differ. This implies that experienced teachers were not able to judge the mean profile of pupils' actual traits more accurately than teacher students and psychology students.

4.2.3.2. Liking. On average, perceivers showed a significantly higher normative accuracy for targets they liked ($b_{21} = 0.13$, $p < .001$). Referred to an individual case, normative accuracy can be interpreted as similarity of the judgment of an individual target to the average target. Consequently, if a teacher liked a specific student, that student was expected to be more similar to the average student. On the other hand, distinctive accuracy had a small, but significant negative relation to liking ($b_{11} = -0.02$, $p < .01$). Or more simply, perceivers' distinctive accuracy was lower for targets they liked.

4.2.3.3. Perceived physical attractiveness. The second relationship moderator, students' physical attractiveness, as judged by the perceivers in each perceiver-target dyad, had a significant positive relationship with normative accuracy ($b_{21} = 0.11$, $p < .001$). Hence, if a teacher evaluated a specific student as attractive, that student was expected to be more similar to the average student in the larger sample. On the other hand, students' attractiveness did not significantly moderate distinctive accuracy ($b_{11} = -0.00$, $p = .88$) suggesting that attractive targets were perceived just as accurately in their distinctive characteristics as less attractive targets.

4.2.3.4. Personality similarity. Perceiver-target personality similarity was overall not significantly related to distinctive accuracy ($b_{11} = 0.01$, $p = .15$) or normative accuracy ($b_{21} = 0.01$, $p = .66$). Given the large range of the personality similarity scores (see Table 3), it can be assumed that this finding cannot be allocated to a low variance of this variable.

Table 6
Fixed effects estimates from the base model with interactions.

Fixed Effects	Distinctive Accuracy			Normative Accuracy		
	Estimate	(*SE*)	*r*	Estimate	(*SE*)	*r*
Liking	−0.02**	(0.01)	.05	0.13***	(0.02)	.17
Attractiveness	−0.00	(0.01)	.00	0.11***	(0.02)	.13
Similarity	0.01	(0.01)	.03	0.01	(0.02)	.01
Teaching experience						
Baseline: Teacher students	0.03	(0.03)	.01	0.72**	(0.06)	.22
Teachers vs. teacher students	−0.01	(0.01)	.01	−0.13*	(0.05)	.05
Psychology students vs. teacher students	−0.01	(0.01)	.01	0.04	(0.05)	.01

Note. ***$p < .001$; **$p < .01$; *$p < .05$. The effect of teaching experience on distinctive and normative accuracy was modeled by computing a dummy variable for experienced teachers and a dummy variable for psychology students. Teacher students served as the baseline group in both cases. Fixed effect estimates are unstandardized regression coefficients; *r* refers to correlations based on a standardized approximation of effect sizes of the estimates ($r = \sqrt{t^2/(t^2 + dfs)}$).

5. Discussion

The present study investigated teacher judgment accuracy across several traits, i.e., profile-based accuracy, by using brief student videos with no prior acquaintance between teachers and students. Moreover, we looked at factors moderating teacher judgment accuracy and demonstrated a first application of SAM as a one-step analytical approach to teacher judgment accuracy.

Our main findings can be summarized as follows: Mean normative accuracy was rather high, with the lowest mean in the teacher group. Mean distinctive accuracy was low and did not differ between teachers, teacher students, and psychology students. Considerable individual differences between the perceivers were found for normative accuracy, but not for distinctive accuracy. Target differences and dyadic variability were moderate and comparable for both normative and distinctive accuracy. Liking and physical attractiveness moderated normative accuracy, whereas liking had a negative effect on distinctive accuracy. Personality similarity was unrelated to both distinctive and normative accuracy.

5.1. Judgment accuracy

On the whole, our findings are in line with results of zero-acquaintance studies in personality and social psychology suggesting that accuracy of teachers' first impressions is comparable to the results of perceivers outside the educational context.

Trait-based accuracies, which were calculated in addition to profile-based accuracies, were significant for each variable in each of the three sample groups, although sample size of targets was small. These results correspond to the findings of Praetorius et al. (2015) who conducted the first zero-acquaintance study with teacher judgements of students' academic self-concept. Results of trait-based accuracies are not directly comparable to results of profile-based accuracies. Trait-based accuracy measured by a correlation coefficient refers to accurate perception of interindividual differences among targets with respect to a *single* trait. Profile-based accuracy, on the other hand, refers to accurate perception of intraindividual differences of a target with respect to *several* traits.

With respect to profile accuracy, our finding is partially consistent with previous research, in which normative accuracy outcomes have shown to regularly outperform distinctive accuracy outcomes (Biesanz, 2010; Human & Biesanz, 2011b, 2012). As normative accuracy outcomes show, teachers' first impressions reflect targets' general pattern of characteristics quite well. Results for distinctive accuracy, on the other hand, show that perceivers on average, do not recognize how the profiles of single targets deviate from the general pattern of characteristics. Contrary to past work in personality and social psychology research, distinctive accuracy did not reach statistical significance. One needs to bear in mind, however, that those studies solely involved judgments of individuals' broad personality traits. Particularly students' academic self-concept and motivation are expected to be considerably more difficult to discern than broad personality traits through thin slices of behavior (Vazire, 2010).

Overall, one can argue that perceiving average students' characteristics may in general be an easier task than perceiving an individual's distinctive set of characteristics. Distinctive accuracy requires detecting and processing an individual's specific characteristics, which depends both on perceivers' ability for cue detection as well as on the amount and quality of cues made available by the targets (Funder, 1995).

Accurate perception of the average students' profile varied to a considerable extent. Considering the differences in normative accuracy (perceptive accuracy) it may be the case that some teachers are prone to a distorted view of the average students' profile and, as a consequence, inappropriate treatment of their class or even age group. As both mean and variability of distinctive accuracy were low, one can conclude that most perceivers were facing difficulties in perceiving students' individual characteristics accurately. For the classroom, this could pose the

risk that single students are misperceived and, as a consequence, are treated inadequately.

Furthermore, some variability was found for targets' judgability with respect to both normative and distinctive accuracy, which indicates that some of the targets may have expressed the respective cues more than others (Funder, 1995). Dyadic variability estimates show that some perceivers judged the profile of certain students particularly accurate but perceived other students' profile particularly inaccurate. In other words, dyadic variability can be seen as an interplay between individual targets and perceivers to result in dyads that are perfectly matched in terms of cue detection (perceivers) and cue expression (targets) (Funder, 1995). In a nutshell, accuracy is not merely a result of an individual teacher's judgmental competencies, but rather based on the interplay between the individual student's expression of valid cues, the individual teacher's susceptibility to such cues and not least the dynamic within the unique judge-target combination (i.e., dyad). Hence, compared to the traditional trait-based approach to teacher judgment research, the applied profile-based approach offers a more comprehensive perspective.

Somewhat unexpected, experienced teachers in the present study had lower normative accuracy outcomes than teacher students. One factor underlying group differences between teachers, teachers students, and psychology students is teaching expertise, which depends, to a considerable extent, on specific knowledge of students resulting from extended periods of teacher-student interaction (Berliner, 2001). Because specific knowledge of students is excluded in a zero-acquaintance design, perceivers can only rely on very general knowledge or even stereotypes. But experienced teachers could actively combat the influence of common stereotypes on their judgments of individual students, and try to do justice to the individuality of each student, although rather unsuccessful as their low distinctive accuracy shows. Moreover, college students, especially teacher students, could be even more familiar than experienced teachers with the average characteristics of students in secondary schools to whom they are closer with regard to age, school experience, and, perhaps, general life-situation. Lacking specific knowledge about students, it is not surprising that experienced teachers did not achieve higher accuracy outcomes than teacher students and psychology students when judging individual differences of targets' actual variable profiles. Although teachers must be considered as professional perceivers, accuracy of their first impressions is not superior to perceivers with less teaching experience.

5.2. Perceiver and relationship moderators and their role for normative and distinctive accuracy

Our finding that liking and students' physical attractiveness were associated with higher normative accuracy is consistent with past work in personality and social psychology research (Human & Biesanz, 2011b; Lorenzo et al., 2010). An explanation for this finding is that liking leads to higher interest and attention towards students, and thereby makes it easier for perceivers to detect cues and process them adequately (Lorenzo et al., 2010).

An alternative interpretation is that students who are perceived as more similar to the average student are liked to a higher degree. Teachers could, in general, like average students more than students deviating from the average because the latter ones could make teaching more challenging or even problematic. Future research should aim at finding out whether or not this assumption can be supported, for instance through interviews with the teachers subsequent to the judgment task. Contrary to liking, the alternative interpretation that targets who are similar to the average are perceived as more physically attractive is not plausible. In our view, the role of cues seems crucial as an explanation of the moderator effects, but alternative interpretations cannot be ruled out completely.

Personality similarity between perceivers and targets showed no significant association with normative accuracy. Unlike in most of the research in personality and social psychology, perceivers and targets in this study were different with respect to many characteristics such as age, status, and general life conditions, and they have different, highly asymmetrical roles. Teachers have to judge students as part of their professional role using categories and constructs that are specific to students. In our study, we consequently used traits that are specific to the school context. In this case, personality similarity might be not very salient for perceivers and, consequently, exert not much influence on their perceptions and their thinking.

In the present study, only liking had a small but significant negative effect on distinctive accuracy. Because variability of distinctive accuracy is low, moderator effects are difficult to interpret. Therefore, we refrain from an interpretation of this small and counterintuitive negative effect that may be due to a special constellation of influences or even random factors.

5.3. Teacher judgments at zero-acquaintance and their relation to classroom-based judgments

Studying judgment accuracy in a zero-acquaintance situation allowed us to purely investigate differences in judgment processes based on the same available information about students and independent from differences in the availability of information (e.g., based on differences in teacher-student interactions and previous knowledge). Perceivers' previous knowledge about the students was not available as reference for the judgments in our study as all students were unacquainted with the perceivers. Instead, the only references for the judgments were students' physical appearance and observable behavior in the brief videos, and this information was held constant across perceivers. The kind of information presented resembles a situation a teacher is regularly confronted with when meeting a student for the first time and forming a first impression about this student, e.g., when taking over a new class.

It is obvious that there are differences between first impressions formed in a zero-acquaintance/thin-slices of behavior situation and a real classroom situation. First, the video presentation has an equal duration of 30 seconds for each target and hence all targets have an equal chance to receive attention. Second, only one target is shown in each video, whereas in the real classroom there is a larger number of students differing with respect to visibility and salience, and consequently each student probably receives a different amount of attention. Third, in the actual classroom the number of available cues is higher and in addition teachers often have prior information about students (e.g., based on student files). It is to date unclear how much and what kind of information is necessary to form first impressions in a real school context. Taking account of these limitations, we believe that a zero-acquaintance/thin-slices approach can be considered as an appropriate tool to study how first impressions are formed in a real classroom situation, at least as a first approximation. But the exact relation between both of these situations has to be explored in more depth.

One of the job requirements of teachers is the professional judgment of their students, for example with respect to grading. But our results suggest that zero-acquaintance judgments of teachers apparently are not profound expert judgments. First, teachers were not able to judge students' individual profile accurately in a zero-acquaintance situation but relied on general knowledge or stereotypes of the average student. Second, their stereotypes were not more accurate than those of teacher students and other college students. Third, normative accuracy was moderated by factors such as liking and physical attractiveness, which are important determinants of social perception in an everyday context. Considering teachers' classroom-based judgments as part of their professional role, it seems somewhat irritating and undesirable that characteristics such as liking and attractiveness would moderate teacher judgments in a professional context. Therefore, part of the requirements that teachers must meet in their professional assessment tasks is to overcome these undesirable influences, which they are usually unaware

of. Teachers should be aware of these biases in order to avoid or control their influence on their professional judgments. Reflecting upon one's own judgments may be an appropriate way to address this problem. Therefore, a practical implication of this research would be to motivate teachers to reflect on their impressions in order to avoid that they would affect their judgments in an unfiltered manner, leading to biased impressions in the classroom. Establishing appropriate forms of teacher training, supervision, and formative evaluation would be helpful to achieve sustainable effects.

5.4. Limitations and further directions

This paper aimed to provide an alternative view on teacher judgment accuracy by combining a zero-acquaintance paradigm with SAM analyses of profile accuracy. Naturally, our first illustration of this approach has some limitations that need to be considered and that give rise to a number of potentially fruitful extensions for future research.

First, participants in the present study were all from one country (i.e., a German university or school), which may limit the generalizability of the obtained results to other countries or geographic regions.

Second, teachers in our sample were teaching different subjects and only a small proportion ($n = 10$) were actual physics teachers. It is unclear, to what extend being familiar to the teaching context would be an advantage also in such early and spontaneous impressions and therefore, future studies on the topic should address teaching subject as a moderator. Moreover, it would be important to replicate the present findings in other subject contexts. Over and above, physics, as much as other STEM disciplines, is a subject that bears a comparatively high risk for judgmental biases, particularly in respect to the gender bias. For future research, it will be therefore interesting to explore the role of students' gender in teachers' first impressions in the STEM subject area.

Third, our analyses did further not include teachers' differing school types and it would be interesting to see if high school teachers, for instance, are having advantages when judging students that belong to the age group they regularly interact with. However, the majority of teachers in our sample were high school teachers, and hence we did not integrate school type as a moderator, which would be a relevant aspect for future research.

An important aspect involves the different number of items that were judged for each construct. In general, there are two strategies for assessment. One of them is item-wise assessment, that is, using the items of the scales applied for measuring the constructs (i.e., criterion variables) for the judgments. As scales for different constructs mostly comprise different numbers of items, this strategy necessarily results in different numbers of judgments for different criteria. Alternatively, in a construct-wise assessment, one item is used for each construct. We applied a mixture of both approaches, using item-wise assessment for most scales but construct-wise assessment for intelligence and academic self-concept. For some constructs such as intelligence, perceivers have a clear intuitive understanding, therefore it seems natural to use construct wise assessment (see also Spinath, 2005). Moreover, as domain-specific constructs can be expected to go along with higher judgment accuracy (trait activation theory; Tett & Guterman, 2000) and the videos showed students working on a physics task, we chose to employ a 1-item scale for students' academic self-concept in physics too. In future research, particular attention should be devoted to these challenges.

Our analyses were confined to profile accuracy. Misperceptions of students' profile, i.e. misperceiving their strengths and weaknesses, are a risk for inappropriate treatment of students. Over- and underestimation of students is another important aspect of biased teacher expectations that is crucial for expectancy effects. Over- and underestimation of students' characteristics is related to the level component of SAM. But it is unclear if the level component is an appropriate indicator of over-/underestimation or only a response bias with respect to the judgment scales. As the level parameter is averaged over several traits, it is questionable if this component reflects a true propensity to overestimate targets' level of traits or merely a response tendency (Cronbach, 1955). Moreover, it is known from self-concept and metacognition research that individuals, especially younger ones, tend to rate themselves more positive than others or than standardized tests would reveal. Therefore, as we are dealing with self-report data on the side of the targets, an interpretation of these findings would not have been straightforward. Thus, for theoretical and methodological reasons it would have been difficult to identify *real* over- and underestimations in the present study. Future studies should consider this question in more detail.

As this study was focused on judgment accuracy as a product of the judgment process, in-depth insights into the process of teachers' impression formation itself should be aimed at in future research. Given the growing evidence of the relevance of the target for the successful judgment process (i.e., *the good target*; Colvin, 1993; Human & Biesanz, 2013), investigating the role of the individual student (e.g., the degree to which a student provides a teacher with valid and relevant cues) may lead to a clearer picture regarding the judgment process as the core aspect of teacher judgment accuracy. In this context, an application of Brunswik's (1956) lens model of perception represents a promising approach (see Back & Nestler, 2016, for an overview in the context of personality judgments). Such an attempt would also extend the present research by integrating areas of moderators that have not been covered in the present work.

Teachers' expectations and acquisition of diagnostic knowledge about their students are assumed to be influenced by first impressions. Whereas accurate first impressions of teachers are often implicitly expected to play an important role in the teaching and learning context, to date, only little evidence exists about the development of those impressions over the course of time and their impact on complex social interactions in the classroom and vice versa. According to the dual-process theory (Evans, 2008; Strack & Deutsch, 2004), those initial judgments, when meeting a student for the first time, could be translated into more deliberate and profound judgments over the course of a school year or even beyond. In this line of thought, two ways of information processing are separated, that is, an implicit, mostly intuitive way of perception and an explicit rather reflected and analytical way of information processing. First impressions are quick and intuitive in nature, which can be regarded as a starting point for a reflected information processing that will eventually and ideally lead to a balanced professional judgment formation. Therefore, various consequences of accurate or erroneous teacher first impressions represent an exciting empirical question that should be investigated applying a longitudinal research design. This approach would also involve the chance to investigate teacher first impressions in a natural classroom context.

6. Conclusion

The present study focused on investigating teacher judgment accuracy using a zero-acquaintance and thin-slice of behavior approach, applying SAM as an analytical strategy for a comprehensive analysis of perceiver and perceiver-target-relationship moderators. Overall, our main findings are in line with essential features of first-impression research on people in general. Our study differs from previous research in personality and social psychology with respect to two main features, namely perceiver-target relationship or role assignment and type of traits used. In personality and social psychology research, perceivers and targets are typically part of the same population, often college students, and traits are chosen that are suitable for peer assessment. For teachers as professionals, moderators such as liking and physical attractiveness have a different meaning than in peer assessment, because these factors can represent biases that impair just and fair assessment of students. If results from zero-acquaintance studies on this topic can be confirmed as stable (Darley & Fazio, 1980; Gunaydin, Selcuk, & Zayas, 2017) and transferable to real classroom situations, such factors need to receive increasing attention in both teacher education and practice.

Our findings indicate the need for further in-depth analyses of teacher-student-relationship moderators in future teacher judgment accuracy research. Compared to traditional teacher judgment accuracy research in actual classroom settings, a zero-acquaintance approach bears the potential to more precisely isolate numerous influences on teacher judgments. It also provides a natural starting point and comparison standard to study more complex ongoing judgment and interaction processes in existing classrooms. It is our hope that the applied zero-acquaintance and thin-slice approach will be further applied and refined in future research, which will help to better understand conditions and processes of teacher judgment accuracy.

Declaration of Competing Interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Acknowledgements

This research project was supported by the German Research Foundation (DFG) through the Graduate School *Teaching and Learning Processes (UPGRADE)* (GRK 1561). The authors thank Katrin Hochdörffer for her support during the data collection.

References

Ambady, Nalini, Hallahan, M., & Rosenthal, R. (1995). On judging and being judged accurately in zero-acquaintance situations. *Journal of Personality and Social Psychology, 69*(3), 518–529. https://doi.org/10.1037/0022-3514.69.3.518.

Ambady, Nalini, & Rosenthal, R. (1992). Thin slices of expressive behavior as predictors of interpersonal consequences: A meta-analysis. *Psychological Bulletin, 111*(2), 256–274. https://doi.org/10.1037/0033-2909.111.2.256.

Anderson, N. H., & Barrios, A. A. (1961). Primacy effects in personality impression formation. *The Journal of Abnormal and Social Psychology, 63*(2), 346–350. https://doi.org/10.1037/h0021966.

Artelt, C. (2016). Teacher Judgments and their Role in the Educational Process. *Emerging Trends in the Social and Behavioral Sciences: An Interdisciplinary, Searchable, and Linkable Resource, 1–16.* https://doi.org/10.1002/9781118900772.etrds040.

Asch, S. E. (1946). Forming impressions of personality. *The Journal of Abnormal and Social Psychology, 41*(3), 258–290. https://doi.org/10.1037/h0055756.

Ashmore, R. D., & Del Boca, F. K. (1981). Conceptual approaches to stereotypes and stereotyping. In D. L. Hamilton (Ed.), *Cognitive processes in stereotyping and intergroup behavior* (pp. 1–31). Lawrence Erlbaum Associates Publishers.

Back, M. D., & Nestler, S. (2016). Accuracy of judging personality. In J. A. Hall, M. Schmid Mast, & T. V. West (Eds.), *The social psychology of perceiving others accurately* (pp. 98–125). Cambridge University Press.

Back, M. D., Schmukle, S. C., & Egloff, B. (2011). A closer look at first sight: Social relations lens model analysis of personality and interpersonal attraction at zero acquaintance. *European Journal of Personality, 25*(3), 225–238. https://doi.org/10.1002/per.790.

Bates, D., Mächler, M., Bolker, B., & Walker, S. (2015). Fitting linear mixed-effects models using lme4. *Journal of Statistical Software, 67*(1), 1–48. https://doi.org/10.18637/jss.v067.i01.

Berliner, D. C. (2001). Learning about and learning from expert teachers. *International Journal of Educational Research, 35*(5), 463–482. https://doi.org/10.1016/S0883-0355(02)00004-6.

Bernieri, F. J., Zuckerman, M., Koestner, R., & Rosenthal, R. (1994). Measuring person perception accuracy: Another look at self-other agreement. *Personality and Social Psychology Bulletin, 20*(4), 367–378. https://doi.org/10.1177/0146167294204004.

Biesanz, J. C. (2010). The social accuracy model of interpersonal perception: Assessing individual differences in perceptive and expressive accuracy. *Multivariate Behavioral Research, 45*(5), 853–885. https://doi.org/10.1080/00273171.2010.519262.

Biesanz, J. C. (2019). The social accuracy model. In Tera D. Letzring, & Jana S. Spain (Eds.), *The Oxford Handbook of Accurate Personality Judgment* (pp. 61–82). New York: Oxford University Press. https://www.oxfordhandbooks.com/view/10.1093/oxfordhb/9780190912529.001.0001/oxfordhb-9780190912529.

Biesanz, J. C., Human, L. J., Paquin, A.-C., Chan, M., Parisotto, K. L., Sarracino, J., & Gillis, R. L. (2011). Do we know when our impressions of others are valid? Evidence for realistic accuracy awareness in first impressions of personality. *Social Psychological and Personality Science, 2*(5), 452–459. https://doi.org/10.1177/1948550610397211.

Borenstein, M., Hedges, L. V., Higgins, J. P. T., & Rothstein, H. R. (2011). *Introduction to meta-analysis.* John Wiley & Sons.

Borkenau, P., & Leising, D. (2016). A more complete picture of personality: What analyses of trait profiles have told us about personality judgment—So far. *Current Directions in Psychological Science, 25*(4), 228–232. https://doi.org/10.1177/0963721416651960.

Borkenau, P., Leising, D., & Fritz, U. (2014). Effects of communication between judges on consensus and accuracy in judgments of people's intelligence. *European Journal of Psychological Assessment, 30*(4), 274–282. https://doi.org/10.1027/1015-5759/a000188.

Brophy, J. E. (1983). Research on the self-fulfilling prophecy and teacher expectations. *Journal of Educational Psychology, 75*(5), 631–661. https://doi.org/10.1037/0022-0663.75.5.631.

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

Brophy, J., & Good, T. L. (1984). Teacher Behavior and Student Achievement. Occasional Paper No. 73. https://eric.ed.gov/?id=ED251422.

Brunswik, E. (1956). *Perception and the representative design of psychological experiments.* University of California Press.

Byrne, D. (1997). An overview (and underview) of research and theory within the attraction paradigm. *Journal of Social and Personal Relationships, 14*(3), 417–431. https://doi.org/10.1177/0265407597143008.

Clifford, M. M., & Walster, E. (1973). The effect of physical attractiveness on teacher expectations. *Sociology of Education, 46*(2), 248–258. https://doi.org/10.2307/2112099.

Cohen, P., Cohen, J., Aiken, L. S., & West, S. G. (1999). The problem of units and the circumstance for POMP. *Multivariate Behavioral Research, 34*(3), 315–346. https://doi.org/10.1207/S15327906MBR3403_2.

Colvin, C. R. (1993). "Judgable" people: Personality, behavior, and competing explanations. *Journal of Personality and Social Psychology, 64*(5), 861–873. https://doi.org/10.1037/0022-3514.64.5.861.

Connelly, B. S., & Ones, D. S. (2010). Another perspective on personality: Meta-analytic integration of observers' accuracy and predictive validity. *Psychological Bulletin, 136*(6), 1092–1122. https://doi.org/10.1037/a0021212.

Corno, L. (2008). On teaching adaptively. *Educational Psychologist, 43*(3), 161–173. https://doi.org/10.1080/00461520802178466.

Cronbach, L. (1955). Processes affecting scores on "understanding of others" and "assumed similarity.". *Psychological Bulletin, 52*(3), 177–193. https://doi.org/10.1037/h0044919.

Darley, J. M., & Fazio, R. H. (1980). Expectancy confirmation processes arising in the social interaction sequence. *American Psychologist, 35*(10), 867–881. https://doi.org/10.1037/0003-066X.35.10.867.

Davis, M. H., & Kraus, L. A. (1997). Personality and empathic accuracy. In W. J. Ickes (Ed.), Empathic accuracy (pp. 144–168). The Guilford Press.

Dusek, J. B., & Joseph, G. (1983). The bases of teacher expectancies: A meta-analysis. *Journal of Educational Psychology, 75*(3), p. https://doi.org/10.1037/0022-0663.75.3.327.

Evans, J. St. B. T. (2008). Dual-processing accounts of reasoning, judgment, and social cognition. *Annual Review of Psychology, 59*(1), 255–278. https://doi.org/10.1146/annurev.psych.59.103006.093629.

Fischer, R., & Milfont, T. L. (2010). Standardization in psychological research. *International Journal of Psychological Research, 3*(1), 88–96. https://doi.org/10.21500/20112084.852.

Frey, A., Taskinen, P. H., Schütte, K., Prenzel, M., Artelt, C., Baumert, J., Blum, W., Hammann, M., Klieme, E., & Pekrun, R. (Eds.). (2009). PISA 2006 Skalenhandbuch. Dokumentation der Erhebungsinstrumente [PISA 2016 scale documentation]. Waxmann.

Funder, D. C. (1995). On the accuracy of personality judgment: A realistic approach. *Psychological Review, 102*(4), 652–670. https://doi.org/10.1037/0033-295X.102.4.652.

Funder, D. C., & Ozer, D. J. (2019). Evaluating effect size in psychological research: Sense and nonsense. *Advances in Methods and Practices in Psychological Science., 2*(2), 156–168. https://doi.org/10.1177/2515245919847202.

Furnham, A., & Monsen, J. (2009). Personality traits and intelligence predict academic school grades. *Learning and Individual Differences, 19*(1), 28–33. https://doi.org/10.1016/j.lindif.2008.02.001.

Gignac, G. E., & Szodorai, E. T. (2016). Effect size guidelines for individual differences researchers. *Personality and Individual Differences, 102*, 74–78. https://doi.org/10.1016/j.paid.2016.06.069.

Gunaydin, G., Selcuk, E., & Zayas, V. (2017). Impressions based on a portrait predict, 1-month later, impressions following a live interaction. *Social Psychological and Personality Science, 8*(1), 36–44. https://doi.org/10.1177/1948550616662123.

Hall, J. A., Andrzejewski, S. A., Murphy, N. A., Mast, M. S., & Feinstein, B. A. (2008). Accuracy of judging others' traits and states: Comparing mean levels across tests. *Journal of Research in Personality, 42*(6), 1476–1489. https://doi.org/10.1016/j.jrp.2008.06.013.

Hall, J. A., Back, M. D., Nestler, S., Frauendorfer, D., Schmid Mast, M., & Ruben, M. A. (2017). How do different ways of measuring individual differences in zero-acquaintance personality judgment accuracy correlate with each other? *Journal of Personality, 86*(2), 220–232. https://doi.org/10.1111/jopy.12307.

Harris, M. J., & Garris, C. P. (2008). You never get a second chance to make a first impression: Behavioral consequences of first impressions. In N. Ambady & J. J. Skowronski (Eds.), First impressions (pp. 147–168). Guilford Publications.

Herppich, S., Wittwer, J., Nückles, M., & Renkl, A. (2013). Does it make a difference? Investigating the assessment accuracy of teacher tutors and student tutors. *The Journal of Experimental Education, 81*(2), 242–260. https://doi.org/10.1080/00220973.2012.699900.

Human, L. J., & Biesanz, J. C. (2013). Targeting the good target: An integrative review of the characteristics and consequences of being accurately perceived. *Personality and Social Psychology Review, 17*(3), 248–272. https://doi.org/10.1177/1088868313495593.

Human, L. J., Biesanz, J. C., Parisotto, K. L., & Dunn, E. W. (2012). Your best self helps reveal your true self: Positive self-presentation leads to more accurate personality impressions. *Social Psychological and Personality Science, 3*(1), 23–30. https://doi.org/10.1177/1948550611407689.

Human, L. J., Sandstrom, G. M., Biesanz, J. C., & Dunn, E. W. (2013). Accurate first impressions leave a lasting impression: The long-term effects of distinctive self-other agreement on relationship development. *Social Psychological and Personality Science, 4*(4), 395–402. https://doi.org/10.1177/1948550612463735.

Human, L. J., & Biesanz, J. C. (2011a). Target adjustment and self-other agreement: Utilizing trait observability to disentangle judgeability and self-knowledge. *Journal of Personality and Social Psychology, 101*(1), 202–216. https://doi.org/10.1037/a0023782.

Human, L. J., & Biesanz, J. C. (2011b). Through the looking glass clearly: Accuracy and assumed similarity in well-adjusted individuals' first impressions. *Journal of Personality and Social Psychology, 100*(2), 349–364. https://doi.org/10.1037/a0021850.

Human, L. J., & Biesanz, J. C. (2012). Accuracy and assumed similarity in first impressions of personality: Differing associations at different levels of analysis. *Journal of Research in Personality, 46*(1), 106–110. https://doi.org/10.1016/j.jrp.2011.10.002.

Human, L. J., Biesanz, J. C., Finseth, S. M., Pierce, B., & Le, M. (2014). To thine own self be true: Psychological adjustment promotes judgeability via personality-behavior congruence. *Journal of Personality and Social Psychology, 106*(2), 286–303. https://doi.org/10.1037/a0034860.

Jussim, L., & Harber, K. D. (2005). Teacher expectations and self-fulfilling prophecies: Knowns and unknowns, resolved and unresolved controversies. *Personality and Social Psychology Review: An Official Journal of the Society for Personality and Social Psychology, 9*(2), 131–155. https://doi.org/10.1207/s15327957pspr0902_3.

Kenny, D. A. (1994). *Interpersonal perception: A social relations analysis*. Guilford Press.

Kenny, D. A., & West, T. V. (2010). Similarity and agreement in self-and other perception: A meta-analysis. *Personality and Social Psychology Review, 14*(2), 196–213. https://doi.org/10.1177/1088868309353414.

Kriegbaum, K., Jansen, M., & Spinath, B. (2015). Motivation: A predictor of PISA's mathematical competence beyond intelligence and prior test achievement. *Learning and Individual Differences, 43*, 140–148. https://doi.org/10.1016/j.lindif.2015.08.026.

Kriegbaum, K., Steinmayr, R., & Spinath, B. (2019). Longitudinal reciprocal effects between teachers' judgments of students' aptitude, students' motivation, and grades in math. *Contemporary Educational Psychology, 59*, 101807. https://doi.org/10.1016/j.cedpsych.2019.101807.

Leising, D., Ostrovski, O., & Zimmermann, J. (2013). "Are We Talking About the Same Person Here?": Interrater agreement in judgments of personality varies dramatically with how much the perceivers like the targets. *Social Psychological and Personality Science, 4*(4), 468–474. https://doi.org/10.1177/1948550612462414.

Letzring, T. D. (2008). The good judge of personality: Characteristics, behaviors, and observer accuracy. *Journal of Research in Personality, 42*(4), 914–932. https://doi.org/10.1016/j.jrp.2007.12.003.

Letzring, T. D. (2010). The effects of judge-target gender and ethnicity similarity on the accuracy of personality judgments. *Social Psychology, 41*(1), 42–51. https://doi.org/10.1027/1864-9335/a000007.

Letzring, T. D. (2015). Observer judgmental accuracy of personality: Benefits related to being a good (normative) judge. *Journal of Research in Personality, 54*, 51–60. https://doi.org/10.1016/j.jrp.2014.05.001.

Letzring, T. D., & Human, L. J. (2014). An examination of information quality as a moderator of accurate personality judgment: Information quality as personality judgment moderator. *Journal of Personality, 82*(5), 440–451. https://doi.org/10.1111/jopy.12075.

Liepmann, D., Beauducel, B., Brocke, B., & Nettelstroh, W. (2012). *IST-Screening. Intelligenz-Struktur-Test*. Hogrefe.

Lipnevich, A. A., & Roberts, R. D. (2012). Noncognitive skills in education: Emerging research and applications in a variety of international contexts. *Learning and Individual Differences, 22*(2), 173–177. https://doi.org/10.1016/j.lindif.2011.11.016.

Lorenz, C. (2011). *Diagnostische Kompetenz von Grundschullehrkräften: Strukturelle Aspekte und Bedingungen*. University of Bamberg Press.

Lorenz, C., & Artelt, C. (2009). Fachspezifität und Stabilität diagnostischer Kompetenz von Grundschullehrkräften in den Fächern Deutsch und Mathematik [Subject-specificity of primary school teachers' diagnistic competence in the subjects German and mathematics]. *Zeitschrift für Pädagogische Psychologie, 23*(34), 211–222. https://doi.org/10.1024/1010-0652.23.34.211.

Lorenzo, G. L., Biesanz, J. C., & Human, L. J. (2010). What is beautiful is good and more accurately understood: Physical attractiveness and accuracy in first impressions of personality. *Psychological Science, 21*(12), 1777–1782. https://doi.org/10.1177/0956797610388048.

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

Marsh, H. W. (1992). Content specificity of relations between academic achievement and academic self-concept. *Journal of Educational Psychology, 84*(1), 35–42. https://doi.org/10.1037/0022-0663.84.1.35.

Möller, J., Pohlmann, B., Köller, O., & Marsh, H. W. (2016). A meta-analytic path analysis of the internal/external frame of reference model of academic achievement and academic self-concept. *Review of Educational Research, 79*(3), 1129–1167. https://doi.org/10.3102/0034654309337522.

Montoya, R. M., & Horton, R. S. (2013). A meta-analytic investigation of the processes underlying the similarity-attraction effect. *Journal of Social and Personal Relationships, 30*(1), 64–94. https://doi.org/10.1177/0265407512452989.

Montoya, R. M., Horton, R. S., & Kirchner, J. (2008). Is actual similarity necessary for attraction? A meta-analysis of actual and perceived similarity. *Journal of Social and Personal Relationships, 25*(6), 889–922. https://doi.org/10.1177/0265407508096700.

Murphy, N. A. (2007). Appearing smart: The impression management of intelligence, person perception accuracy, and behavior in social interaction. *Personality and Social Psychology Bulletin, 33*(3), 325–339. https://doi.org/10.1177/0146167206294871.

Murphy, N. A., Hall, J. A., & Colvin, C. R. (2003). Accurate intelligence assessments in social interactions: Mediators and gender effects. *Journal of Personality, 71*(3), 465–493. https://doi.org/10.1111/1467-6494.7103008.

Nickerson, R. S. (1998). Confirmation bias: A ubiquitous phenomenon in many guises. *Review of General Psychology, 2*(2), 175–220. https://doi.org/10.1037/1089-2680.2.2.175.

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*(6), 936–943. https://doi.org/10.1177/0021934705285955.

Praetorius, A.-K., Drexler, K., Rösch, L., Christophel, E., Heyne, N., Scheunpflug, A., … Dresel, M. (2015). Judging students' self-concepts within 30s? Investigating judgement accuracy in a zero-acquaintance situation. *Learning and Individual Differences, 37*, 231–236. https://doi.org/10.1016/j.lindif.2014.11.015.

Praetorius, A.-K., Karst, K., Dickhäuser, O., & Lipowsky, F. (2011). Wie gut schätzen Lehrkräfte die Fähigkeitsselbstkonzepte ihrer Schüler ein? [How accurate are teachers when judging students' self-concepts?]. *Psychologie in Erziehung und Unterricht, 2*, 81–91.

R Development Team. (2020). *R: A language and environment for statistical computing*. R Foundation for Statistical Computing. https://www.R-project.org/.

Rammstedt, B., & John, O. P. (2007). Measuring personality in one minute or less: A 10-item short version of the Big Five Inventory in English and German. *Journal of Research in Personality, 41*(1), 203–212. https://doi.org/10.1016/j.jrp.2006.02.001.

Ready, D. D., & Wright, D. L. (2011). Accuracy and inaccuracy in teachers' perceptions of young children's cognitive abilities: The role of child background and classroom context. *American Educational Research Journal, 48*(2), 335–360. https://doi.org/10.3102/0002831210374874.

Reynolds, D. J., & Gifford, R. (2001). The sounds and sights of intelligence: A lens model channel analysis. *Personality and Social Psychology Bulletin, 27*(2), 187–200. https://doi.org/10.1177/0146167201272005.

Ritts, V., Patterson, M. L., & Tubbs, M. E. (1992). Expectations, impressions, and judgments of physically attractive students: A review. *Review of Educational Research, 62*(4), 413–426. https://doi.org/10.3102/00346543062004413.

Rost, D. A., Sparfeldt, J. R., & Schilling, S. R. (2007). DISK-GITTER mit SKSLF-8. Differentielles Schulisches Selbstkonzept-Gitter mit Skala zur Erfassung des Selbstkonzepts schulischer Leistungen und Fähigkeiten. [Differential academic self-concept-scale including a scale to assess the self-concept regarding academic achievement and abilities]. Hogrefe.

Schunk, D. H., & Zimmerman, B. J. (2012). *Motivation and self-regulated learning: Theory, research, and applications*. Routledge.

Seidel, T., Prenzel, M., Duit, R., & Lehrke, M. (2003). Technischer Bericht zur Videostudie "Lehr-Lern-Prozesse im Physikunterricht" [Technical report following the video study "Teaching and learning processes in the physics classroom"]. IPN.

Skinner, E. A., & Belmont, M. J. (1993). Motivation in the classroom: Reciprocal effects of teacher behavior and student engagement across the school year. *Journal of Educational Psychology, 85*(4), 571–581. https://doi.org/10.1037/0022-0663.85.4.571.

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

Strack, F., & Deutsch, R. (2004). Reflective and impulsive determinants of social behavior. *Personality and Social Psychology Review, 8*(3), 220–247. https://doi.org/10.1207/s15327957pspr0803_1.

Südkamp, A., Kaiser, J., & Möller, J. (2012). Accuracy of teachers' judgments of students' academic achievement: A meta-analysis. *Journal of Educational Psychology, 104*(3), 743–762. https://doi.org/10.1037/a0027627.

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

Tett, R. P., & Guterman, H. A. (2000). Situation trait relevance, trait expression, and cross-situational consistency: Testing a principle of trait activation. *Journal of Research in Personality, 34*(4), 397–423. https://doi.org/10.1006/jrpe.2000.2292.

Urhahne, D., Chao, S.-H., Florineth, M. L., Luttenberger, S., & Paechter, M. (2011). Academic self-concept, learning motivation, and test anxiety of the underestimated student. *British Journal of Educational Psychology, 81*(1), 161–177. https://doi.org/10.1348/000709910X504500.

Vazire, S. (2010). Who knows what about a person? The self–other knowledge asymmetry (SOKA) model. *Journal of Personality and Social Psychology, 98*(2), 281–300. https://doi.org/10.1037/a0017908.

Wason, P. C. (1960). On the failure to eliminate hypotheses in a conceptual task. *Quarterly Journal of Experimental Psychology, 12*(3), 129–140. https://doi.org/10.1080/17470216008416717.

MANUSCRIPT 3

Appearing Smart, Confident and Motivated: A Lens Model Approach to Teacher Judgment Accuracy

Caroline V. Bhowmik[1], Mitja D. Back[2], Steffen Nestler[2], & Friedrich-Wilhelm Schrader[1]

[1]Department of Psychology, University of Kaiserslautern-Landau, Germany

[2]Department of Psychology, University of Münster, Germany

Abstract

Which behavioral and visual information do teachers rely on when judging relevant characteristics of their students and which cues should they rely on? Drawing on Brunswik's (1956) lens model we investigated the role of students' expression of nonverbal behavioral cues (e.g., friendly facial expression) and physical appearance (e.g., wearing eyeglasses) and how this information is utilized during the judgment process by pre-service teachers and psychology students ($N = 102$). Perceivers provided ratings of students' ($N = 45$) academic self-concept, intelligence and motivation in brief nonverbal video clips showing one student each in a physics classroom. Numerous behavioral and physical cues (in total 165) were extracted from the stimulus material by two independent raters. Perceivers achieved highest accuracy for students' motivation ($r = .21$ for single-perceiver and $r = .43$ for average-perceiver accuracy) and intelligence was judged with the lowest accuracy ($r = .05$ for single-perceiver and $r = .10$ for average-perceiver accuracy). Lens model parameter analysis indicated that perceivers strongly relied on students' sex, an attentive and self-assured facial expression, and whether or not a student was wearing eyeglasses in their judgments. Valid cues, on the other hand, involved students' sex, a masculine and distinctive appearance, and a tensed as well as friendly facial expression. An overall favorable judgment for boys points into the direction of a gender bias. Implications for our understanding of teacher judgment processes and outcomes are discussed.

Keywords: Teacher judgment accuracy, Brunswik's lens model, diagnostic competence, thin-slices of behavior, self-concept, intrinsic motivation

1 Introduction

Teachers' judgments about students are ubiquitous phenomena in the school context and they pertain to a broad range of characteristics (e.g., abilities, self-concept, motivation). These judgments can have an impact on teachers' decisions and behavior, as well as on teacher-student interaction and students' perceptions, motivation, and behavior. Accurate teacher judgments and impressions are important for everyday instruction, such as adjusting the teaching content and instruction to students' aptitudes and prior knowledge (*adaptive teaching*; e.g., Hardy et al., 2019). Inaccurate or biased teacher perceptions on the other hand can have unfavorable effects on students' achievement, motivation and even life satisfaction (*expectancy effects*; e.g., Bergold, 2023; Friedrich et al., 2015; Jussim & Harber, 2005). Teacher judgment accuracy is typically measured by comparing teachers' judgments of students' characteristics with students' actual characteristics (criteria) measured by tests or some kind of self-report data. Traditionally, teachers' judgment accuracy is assessed at long-time acquaintance, i.e., for students whom the teacher knows for an extended period of time. In contrast, teachers' judgments at zero-acquaintance, which can be investigated on the basis of short behavioral episodes (zero acquaintance and thin-slices-of-behavior approach, Ambady et al., 1995; Ambady & Rosenthal, 1992), are to date poorly explored. Zero-acquaintance judgments in the school context refer to situations when a teacher meets a student for the first time and forms a first impression of the student. These first impressions are important because given that they are often stable (Darley & Fazio, 1980; Harris & Garris, 2008; Nickerson, 1998) they can determine subsequent judgments and long-term evaluations. Moreover, initial judgments can influence teachers' instructional and managment decisions when teaching a school class for the first time, for example by identifying students who may rather support or disrupt a planned teaching sequence. There is a lack of knowledge about factors influencing teacher's judgment accuracy in a zero-acquaintance situation as well as about the judgment process per se, including the cognitive and behavioral aspects that lead to certain judgments (Schnitzler et al., 2020). In this study, we use Brunswik's Lens Model (BLM; Brunswik, 1956) to analyze factors that contribute to judgment accuracy with respect to important student characteristics (intelligence, self-concept and motivation).

1.1 Perceiver Judgments and Judgment Accuracy at Zero-Acquaintance

Most frequently, accuracy is measured by correlating an individual teacher's judgments and the actual characteristics of the students in his or her class (i.e., *rank order accuracy*, also referred to as *rank component*; Helmke & Schrader, 1987; Praetorius et al., 2015; Südkamp et al., 2012). For traditional studies that serve as a reference, meta-analyses indicate an average accuracy of r = .63 for students' academic achievement (Südkamp et al., 2012), and r = .50 for cognitive ability (Machts et al., 2016). Accuracies of non-cognitive student characteristics are usually lower ($.29 \leq r \leq .55$ for academic self-concept and $.10 \leq r \leq .20$ for school anxiety and motivation; Praetorius et al., 2011, 2015; Spinath, 2005; Urhahne et al., 2011).

Research in social and personality psychology suggests that a *thin-slices of behavior* and *zero-acquaintance* (Ambady et al., 1995; Ambady & Rosenthal, 1992) approach could be an important extension of traditional accuracy research in the context of education. Studies in which perceivers do not know the targets (zero-acquaintance) and observation is restricted to short behavioral episodes (thin-slice of behavior) could hence be a promising tool to study teachers' first impressions of students' personality characteristics which are relevant to learning (Bhowmik et al., 2021).

Previous research in social and personality psychology using a zero-acquaintance or thin-slices-of-behavior approach has shown that personality characteristics, such as the Big Five, and other characteristics, such as intelligence, are usually perceived with substantial accuracy (Ambady & Rosenthal, 1992; Borkenau et. al, 2014; Murphy, 2007; Murphy et al., 2003; Reynolds & Gifford, 2001). In this research, judgments are mainly based on indicators related to the physical appearance and general behavioral pattern of a target, e.g., gestures. The application of brief videos as stimulus material for the perceivers is based on a large amount of research showing that even very short glimpses of expressive behavior can already provide valid information about social variables, such as personality or emotion (Ambady & Rosenthal, 1992; Murphy et al., 2015, 2019).

So far, only few studies addressed teacher judgments and judgment accuracy based on minimal information. In a study by Praetorius et al. (2015), teacher judgment accuracy regarding students' academic self-concept was investigated based on 30-second videos of students and no prior acquaintance between teachers and students. The average accuracy values in the four zero-acquaintance samples ranged between r = .31 and r = .39, which corresponds roughly to the average accuracy achieved in a natural classroom sample (r = .29, SD = .34). This result shows that teachers

were indeed able to judge the rank order of unacquainted students to some degree and, remarkably, knowing a student well did not increase teachers' accuracy outcomes. In a study by Lansu and Berg (2020), different groups of perceivers (teachers, students, and young adults) judged students' likeability, popularity, prosocial behaviour, aggression, and level of exclusion based on brief 20-second long videos. The researchers found better than chance accuracy outcomes based on thin-slices of behavior for students' popularity (r = .16) and prosocial behavior (r = .21), but not for students' aggression (r = -.14) and level of exclusion (r = -.09). Similar to the findings by Praetorius et al. (2015), familiarity with the social context of the targets did not benefit judgment accuracy. In another recent zero-acquaintance study (Bhowmik et al., 2021) small to moderate rank order accuracies were obtained for the broad academic self-concept, domain-specific self-concept and intrinsic motivation in physics, and intelligence.

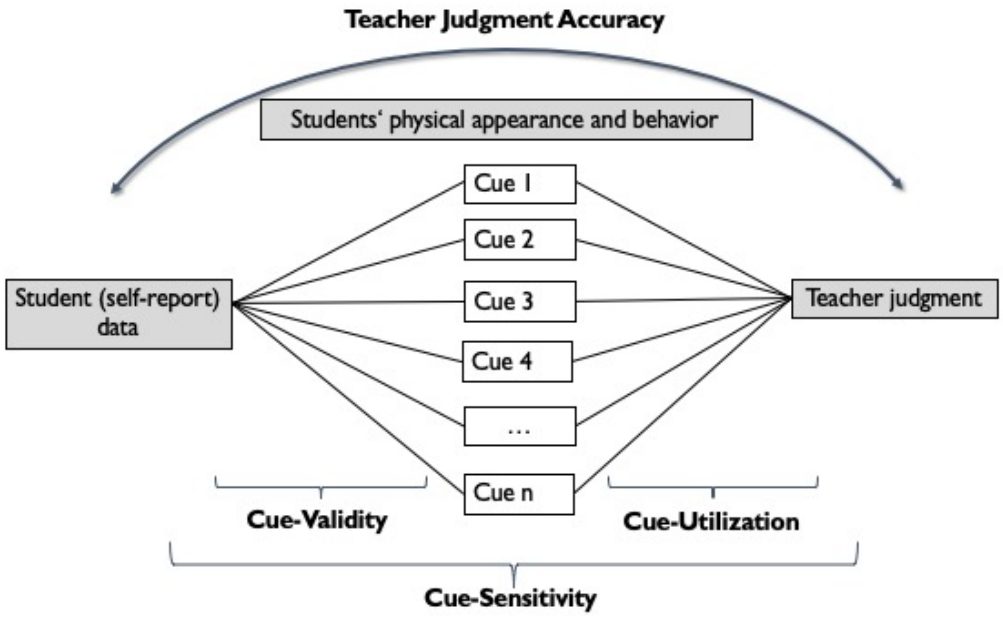
1.2 Judgment Accuracy within the BLM Framework

Previous studies on teacher judgment accuracy did only rarely examine the information judgements are based on. Thus, it remains an open question how teachers obtain a more or less accurate judgment. Brunswik's (1956) Lens Model (BLM) offers a theoretical and methodological framework to identify cues teachers rely on when rating students' individual characteristics. It is a conceptual model that has been extensively employed to study the processes that are involved in interpersonal perception and that can lead to either high or low judgment accuracy (see Figure 1 for a graphical illustration of the model and Back & Nestler, 2016; Nestler & Back, 2013). The basic idea of the lens model is that in order to infer other individuals' personality dispositions or inner states, such as their motivation or emotion, which are not directly observable, a perceiver has to use observable information (cues; see the middle part of Figure 1) that is related to these dispositions or inner states (i.e., the criterion to judge). Whether or not certain cues are related to the criterion, that is, the actual value of the targets' characteristics, such as for instance students' self-reported motivation, is referred to as *cue validity.* The extent to which a given set of cues and hence the amount of visible information is generally able to predict the targets' criterion values is called *predictability.* The strength with which perceivers utilize a certain cue for their judgment is referred to as *cue utilization* and the extent to which a perceiver applies the same judgment strategy consistently, i.e., utilizes the same set of cues with equal strength across all targets, is called *response consistency*. Finally, the extent to which a perceiver utilizes valid cues is called *cue*

sensitivity or *matching*, the parameter of a lens model that describes how well a perceiver's judgment model fits with the model of the criterion.

Figure 1

Diagram of a Lens Model to Describe Personality Judgments



Note. The figure was adapted from "Applications and Extensions of the Lens Model to Understand Interpersonal Judgments at Zero Acquaintance" by S. Nestler und M. D. Back, 2013, *Current Directions in Psychological Science*, *22*(5), p. 375

Drawing on the basic conjecture of BLM, perceptual accuracy is more likely to occur the more valid cues exist in the judgment context (predictability) and the more sensitive a perceiver is in respect to identifying the validity of individual cues (i.e., the more valid cues she or he uses; cue sensitivity). Ideally, a perceiver then applies his or her cue knowledge consistently across targets and situations (response consistency) (Nestler & Back, 2013).

Previous research applying BLM has shown that cues extracted from thin-slices of behavior do not only predict judgments but are also related to the actual target characteristics (Back et al., 2011; see Breil et al., 2021 for an overview; Karelaia & Hogarth, 2008; Reynolds & Gifford, 2001). However, so far only few studies have applied the lens model in teacher judgment accuracy research (Förster & Böhmer, 2017). In an early study, Cooksey et al. (1986) investigated cues that

teachers utilize when judging reading comprehension of kindergarten children. Whereas teacher students arrived at accurate judgments on average, they showed differences in the validity of the information they used in the judgment process. In a more recent study, Marksteiner et al. (2012) explored cues that pre-service teachers use to identify cheating, operationalized via true vs. invented stories. Results show that the cues reported by the pre-service teachers predicted their judgments, but were not predictive of the objective cheating status (Marksteiner et al., 2012). In a study by Schnitzler et al. (2020), pre-service teachers rated student profiles (i.e., strong, struggling, overestimating, underestimating, uninterested) using 11-minute long videos of a classroom situation. Results showed that judgment accuracy was related to the utilization of specific combinations of cues. However, information regarding cue utilization was based on pre-service teachers' self-reports. Thus, these results only show which cues people *think* they used, but not which cues they actually used. To answer this question, one has to assess all cues that may be relevant in a way that is independent from the perceivers.

2 The Present Research

In this study, we explored nonverbal behavioral cues (e.g., gestures, facial expressions) and physical cues (e.g., clothing style, eyeglasses) perceivers use when judging unknown students' intelligence, self-concept, and motivation after a short period of observation. These characteristics were chosen, because they are particularly important for teacher-student interactions as well as students' academic outcomes (Kriegbaum et al., 2015; Machts et al., 2016; Möller et al., 2016). Intelligence as it is applied in this study refers to the ability to infer relations and regularities (Liepmann et al., 2012). General academic self-concept (Rost et al., 2007) describes a student's overall positive or negative evaluation of his or her school performance, whereas the subject-specific academic self-concept in physics (Seidel et al., 2003) is defined as a student's perception of his or her ability in physics. Intrinsic motivation in physics (Frey et al., 2009) refers to the enjoyment of activities in physics and physics lessons.

Perceivers' (teacher and psychology students in the role of teachers) judged students shown in brief video clips with respect to the aforementioned characteristics. To identify cues describing students' appearance and behavior, a coding manual that has been applied in earlier research (Back et al., 2011; Nestler et al., 2012) was used and adapted to cover all relevant cue domains. To obtain cue values, the videos were rated independently by two trained raters. In this study, only appearance

and nonverbal cues were used but no verbal cues as in a first step the aim was to find out if basic visual information is sufficient to predict judgments and criteria.

In order to provide an opportunity to observe a great amount of highly visible cues, we chose a setup in which a student (target) works independently on a physics experiment. Doing experiments results in much more observable behaviors than seatwork or group discussions. Given that the experimental situation was identical for each target, we ensured a constant research setting in which variability was mostly due to differences between the targets themselves.

By applying a zero-acquaintance and thin-slice of behavior approach, we moreover aimed at capturing some central features of first impression formation that translate to actual situations in which teachers are confronted with previously unknown students for the first time. Investigating teacher judgments with no prior acquaintance between teachers and students bears the potential to disentangle the judgment process per se, independent of previous teacher-student interactions and knowledge or information about the students.

To examine accuracy and its components, we followed the BLM logic[1] with the aim to investigate the following four questions:

(1) How accurate are perceivers' judgments of students' academic self-concept, intrinsic motivation and intelligence based on brief videos and no prior acquaintance?
(2) Does the environment contain information that is relevant to the judgment task (cue validity and predictability)? Does the validity of the individual cues differ across the four investigated characteristics? (3) Does the perceiver utilize the given information in his or her judgment and does he or she use this particular information in the same manner across targets and situations (cue utilization and response consistency)? Which cues does the perceiver utilize and are there differences in cue utilization across the four investigated characteristics? (4) Is the information a perceiver utilizes for his or her judgments valid (cue sensitivity)?

[1] This understanding of accuracy is also reflected in the lens model equation: $r_a = r_m R_e R_s$, where r_m (matching index) is the correlation between the predicted values of both the judgment and the criterion model, R_e refers to predictability and R_s refers to consistency, which is measured by the correlation between the actual judgments and the judgments predicted by the cues (see Karelaia & Hogarth, 2008, for a detailed elaboration of the lens model equation).

3 Method

3.1 Procedure and Design

Perceivers watched brief videos (i.e., 45 seconds length), showing one of 45 students (target subjects) each while working on a physics experiment task during a school visit at the university's physics laboratory. The data collection was conducted through a computer-based procedure: Each perceiver received a laptop and after a brief trial round with two videos, which were excluded from the later analyses, the actual judgment task was carried out. Perceivers were then shown the video snippets of all 45 targets and were prompted by the program after each video to provide their judgments regarding targets' intelligence, academic self-concept in physics, intrinsic motivation in physics, and the broad academic self-concept. To control for sequence effects, all videos, as well as the appearance of the criterion variables, were randomized. Moreover, all videos were muted before they were used for data collection so that audible speech could not serve as a cue in our study.

The criterion data – both the targets' self-reports and the intelligence test – was assessed during a school visit of the research team prior to visiting the university. Moreover, demographic information of the perceivers was collected during a separate day prior to the actual judgment task.

3.2 Target Subjects

Students ($n = 45$, 42% female, 14 to 17 years, $M_{age} = 15.6$, $SD_{age} = .68$) in tenth grade served as targets. They were part of a larger group of 10th grade students ($N = 244$) from ten secondary schools in Germany, who were visiting the university's physics laboratory. The students were video recorded during their visit in the laboratory while independently working on physics experiments in dyads. From the resulting video material, 45 nonverbal, brief videos (45 seconds each) were extracted showing one student each. When selecting the video snippets for the study, we aimed at ensuring that differences between targets in the criterion variables are sufficiently large. Therefore, students in the final stimulus material displayed a high variance of intelligence, motivation and academic self-concept. During the video selection process, we also ensured that the video snippets are comparable and of the same quality (i.e., mostly frontal gazing student, no interaction with other students and/or the teacher and a comparable resolution and light). Informed consent of both the participating students and parents was obtained prior to the conduction of this study.

3.3 Perceiver Subjects

In total, N = 102 undergraduate students (72 % female, 18 to 32 years, M_{age} = 23.05, SD_{age} = 2.63) enrolled in the teacher (n = 78; 68% female, 18 to 31 years, M_{age} = 23.29, SD_{age} = 2.49) and psychology (n = 24; 83% female, 19 to 32 years, M_{age} = 22.25, SD_{age} = 2.94) study programs at the University of Koblenz-Landau, Germany, served as perceivers in our study. They received course credit (psychology students) or a gift voucher (pre-service teachers) in exchange for their participation. All perceivers were previously unacquainted to the students.

3.4 Criterion Measures

Targets' *general academic self-concept* was assessed by using the *DISK-Gitter* (Rost et al., 2007) containing eight items (e.g., "I have a good feeling regarding my performance at school") with categories ranging from 1 (*do not agree*) to 6 (*fully agree*). Targets' *academic self-concept in physics* was measured using a four-item scale (e.g., "I am talented for physics"; Seidel et al., 2003). Answering categories ranged from 1 (*do not agree*) to 4 (*fully agree*). Targets' *intrinsic motivation* for physics was measured using the PISA 2006 (Frey et al., 2009) scale (e.g. "I enjoy attending physics lessons in school") comprising three items. Answering categories ranged from 1 (*never*) to 4 (*almost always*). Finally, targets' *intelligence* was measured by the IST-Screening (Liepmann et al., 2012). This test consisted of three subtests: word analogies, numerical series, and matrices. Raw scores were transformed into IQ-Scores based on the test manual of the IST-Screening.

3.5 Perceiver Judgments

Students' *general academic self-concept* and the *academic self-concept in physics* were each rated based on three items from the original scales. For the general self-concept (e.g., "The student has a good feeling about his/her engagement in school") the rating categories ranged from 1 (*do not agree*) to 6 (*fully agree*) and from 1 (*do not agree*) to 4 (*fully agree*) for the academic self-concept in physics (e.g., "The student thinks he or she is gifted for physics"). Students' *intrinsic motivation* in physics was rated on two items from the three-item scale that was applied in assessing students' self-reports (e.g., "The student enjoys attending physics lessons in school") with the 4 - point scale ranging from 1 (*fully disagree*) to 4 (*fully agree*). The final items that were used for the perceiver judgments were selected based on internal consistency measures. Students'

intelligence was rated by using a 1-item scale ("In your opinion, how intelligent is this student?"). Judgment categories ranged from 1 (*not intelligent*) to 6 (*very intelligent*).

3.6 Cue Measures

The selection of cues was based on evidence from earlier research applying BLM (Back et al., 2011; Breil et al., 2021; Nestler et al., 2012; Stopfer et al., 2014). Following this, a coding manual was developed with the aim to cover all possibly relevant cue areas. The manual therefore included numerous static and dynamic cues to depict as much information from the brief videos as possible. Whereas static cues described the appearance of a student, such as body size, hair color or whether he or she was wearing eyeglasses, dynamic cues described students' facial expression (i.e., smiling or attentive facial expression) and gestures (i.e., tensed or expressive gestures). Both facial expression and gesture cues were defined considering essential psychological dimensions: expressivity vs. introversion, negative affection, aggressivity and arrogance, agreeableness and warmth, dominance and self-assuredness, and motivation and attentiveness. Essentially, following BLM standards, we intended to extract everything visible that could be potentially relevant to predict either the actual student characteristic or the judgments of the perceiver. Based on the coding manual, the video material (N = 45 videos/targets) was coded by two trained and independent raters resulting in a targets (rows) x cues (columns) data matrix. The raters were familiarized with the coding manual prior to the actual rating process and were asked to code two trial video snippets, both of which were not included in the later analyses. Students' sex, a static cue variable, was coded with 0 for boys and 1 for girls. Given that each of the two raters coded all of the 45 videos, Cronbach's Alpha could be calculated as indicator for the inter-rater reliability, which can overall be considered high with α =.84 for nonverbal (i.e., dynamic) and α =.89 for physical (i.e., static) cues. All cues were then z-standardized within targets' sex to ensure that mean differences within the cues between the boys and girls in the sample are excluded. As many variables (cues) were statistically related within as well as across dimensions, we attempted to reduce the complexity and number of available cues by combining some of them into larger cue aggregates based on theoretical considerations (see Breil et al., 2021; Nestler et al., 2012; Stopfer et al., 2014 for similar approaches).

The calculations resulted in 17 final cues, of which 7 are cue aggregates (see Table 1 for an overview of the cues and resulting aggregates). The intercorrelations between the final cues are displayed in the Appendix.

Table 1

Final Cue Categories and Internal Consistencies (α) of Cue Aggregates

Cue category (α)	Cues before aggregation and single cue measures	Cue category (α)	Cues before aggregation and single cue measures
Static cues		Dynamic cues	
Attractiveness (.88)	Hair: tidy, tied together, attractive, straight Body: slim, tidy Face: attractive, clear Clothing: tidy, attractive, modern	Facial expression	
		Friendly (.96)	Number, duration, warmth and intensity of smiling Friendly facial expression
Masculinity (.86)	Masculine clothes, hair, face, body, absence of make-up, shortness of hair	Self-assured, non-shy (.65)	Non-shy gaze Self-assured facial expression
Dark appearance (.65)	Dark face Dark hair	Communi-cative	Single cue (gaze to interaction partner)
Distinctive clothing style (.72)	Distinctive clothing Colourful clothing	Attentive	Single cue
Distinctive hair	Single cue	Gestures	
Distinctive face	Single cue		Fast gestures
Mature face	Single cue		Active movements
Body size	Single cue	Expressive (.95)	Expressive hand gestures
Apparent social background	Single cue		Insulting gestures
Eyeglasses	Single cue	Tensed	Single cue
Sex (boys: 0, girls: 1)	Single cue		

Note. Cronbach's Alpha is defined only for cue aggregates.

3.7 Lens Model Analyses

All analytical procedures described below were conducted for each student characteristic separately.

3.7.1 Judgment Accuracy

Judgment accuracy is measured by the correlation between perceiver judgments of the targets and the targets' criterion values, calculated over targets. *Single-perceiver accuracy* refers

to the average correlation between a single perceiver's judgments of the different targets and the criterion values of these targets. To compute the average of single-perceiver accuracies, single-perceiver correlations were Fisher- *z* transformed, averaged and transformed back into a correlation. *Aggregate-perceiver accuracy* (Back & Nestler, 2016) is the correlation between average perceiver judgments of the targets and targets' criterion values. While average values of single-perceiver accuracy characterize the mean accuracy of individual perceivers, aggregate-perceiver accuracy characterizes the achievement of a hypothetical average perceiver.

3.7.2 Cue Utilization

Cue utilization is obtained by regressing a single perceiver's judgments of the targets as dependent variable on the cue values as independent variables (i.e., single-perceiver utilization). The regression weight of a cue indicates the utilization of this specific cue for the specific perceiver. By contrast, average-perceiver cue-utilization refers to the regression model of the average perceiver judgments on the aggregated cues. In addition to the regression models, correlation coefficients were calculated accordingly to obtain cue utilization values for a single and an average perceiver.

3.7.3 Response Consistency

A single perceiver's *response consistency* is measured by the correlation between the actual judgments and the judgments predicted by the cues. It is a measure of how good perceivers' judgments are approximated by a linear model or decision rule. Response consistency would be perfect when a perceiver would use the same linear decision rule, i.e., equal weights for each target. Varying weights for the same cue for different targets would result in a departure from a general linear decision rule and a less-than-perfect linear prediction.

3.7.4 Cue Validity and Predictability

Cue validity describes the extent to which a cue predicts the criterion. To obtain cue validities, multiple regressions were computed in which the targets' criterion values of a construct were regressed on the respective cues. The resulting regression weights are indicators of the cue validities. The regression model characterizes the model of the environment for a specific criterion. *Predictability* is the correlation between the predicted value of the model with the actual target

values, indicating the degree to which the cues in total predict the criterion. In addition to the regression analysis carried out to determine cue validity, we also computed correlation-based cue validity by correlating targets' criterion values with the cue values of the 17 cues.

3.7.5 Cue Sensitivity

Cue sensitivity or *matching* refers to the relation between the model of the perceiver and the model of the environment. Matching scores were calculated by computing the correlation between the predicted values of the regression model in which a perceiver's judgments of a respective criterion were regressed on the cues and the predicted values of the regression model in which targets actual values were regressed on the cues. Matching is a measure of cue sensitivity as it shows how much the utilization of cues by the perceiver correspond to their validity (i.e., how well perceivers use cues according to their validity).

In all regression analyses, static and dynamic cues were integrated in the same model. All analyses were carried out with the statistical software *R*, version 4.0.0 (*R* Development Core Team, 2020) and the packages *plyr* (Wickham, 2022)*, psych* (Revelle, 2023) and *quantpsych* (Fletcher, 2022).

4 Results

Descriptive statistics of students' criterion values and perceiver judgments are described in Table 2. Table 3 presents results obtained from the lens model parameter analyses, including single-perceiver and average-perceiver accuracy, consistency, sensitivity and predictability coefficients for all assessed constructs.

4.1 Judgment Accuracy

How accurately did perceivers judge targets' general and domain-specific academic self-concept, intrinsic motivation, and intelligence based on brief, nonverbal video clips? Mean accuracies were significant for all four students' characteristics, although mean accuracy scores were only low to medium[2]. On average, single perceivers were not successful at

[2] To evaluate effect sizes obtained in this study, we employed the benchmarks suggested by Funder & Ozer (2019) as orientational framework.

Table 2

Means (M), Standard Deviations (SD), Minimum (Min), Maximum (Max), Internal Consistencies (α) and Intercorrelations of Targets' Characteristics (upper line) and Perceiver Judgments (lower line)

Criterion variables	*M*	*SD*	*Min*	*Max*	*α*	1	2	3
1. Academic self-concept								
Targets characteristics	4.00	0.82	2.50	5.75	.80	-		
Perceiver judgments	3.34	0.37	2.65	4.34	-	-		
2. Academic self-concept in physics								
Targets characteristics	2.68	0.77	1.00	4.00	.85	**.64**	-	
Perceiver judgments	2.47	0.19	1.99	2.99	-	**.59**	-	
3. Intrinsic motivation physics								
Targets characteristics	2.76	0.87	1.33	4.00	.89	**.52**	**.75**	-
Perceiver judgments	2.41	0.28	1.78	5.16	-	**.49**	**.72**	-
4. Intelligence								
Targets characteristics	108.94	12.77	82.00	134.50	-	**.15**	**.28**	**.10**
Perceiver judgments	3.89	0.45	2.93	5.16	-	**.42**	**.38**	**.29**

Note. N_{Target} = 45; $N_{Perceiver}$ = 102. Internal consistencies (α) refer to the complete student sample (N = 244). Intelligence: IQ-scale and reliabilities (Cronbach's α, split half, test-retest between .72 and .90) based on test validation studies for the IST-Screening. Bold correlations p ≤ .05 (t-test, two-tailed).

Table 3

Single-Perceiver Accuracy (SAcc), Average-Perceiver Accuracy (AAcc), Cue Sensitivity (CS), Response Consistency (RC) and Predictability (PR)

Criterion variables	SAcc	AAcc	Variance $_{SAcc}$	RC	Variance $_{RC}$	CS	Variance $_{CS}$	PR (R^2)
1. Self-concept	**.05**	**.10**	0.02	**.69**	0.02	.20	0.05	.48
2. Self-concept physics	**.13**	**.24**	0.02	**.72**	0.03	.26	0.04	.49
3. Intrinsic motivation	**.23**	**.43**	0.01	**.72**	0.02	**.37**	0.05	**.61**
4. Intelligence	**-.03**	**.07**	0.02	**.70**	0.02	-.09	0.05	.29

Note. $N_{Perceiver}$ = 102; N_{Target} = 45. SAcc were Fisher-z transformed, averaged and transformed back into correlations. Bold correlations $p \leq .05$ (t-test; two-tailed).

judging students' general academic self-concept (r = .05) and intelligence (r = -.03), but could to some extent rate students' academic self-concept in physics (r = .13) and intrinsic motivation in physics (r = .23) accurately. In contrast to individual perceivers (SAcc), accuracies of the average perceiver (AAcc) were higher, ranging from r = .10 for intelligence to r = .43 for intrinsic motivation in physics.

4.2 Predictability, Response Consistency, and Cue Sensitivity

The degree to which the employed set of cues was suitable to predict targets' characteristics differed widely across the four selected judgment criteria. Hence, *predictability* was most evident for students' intrinsic motivation in physics (R^2 = .61), but for students' intelligence predictability was rather low (R^2 = .29). For students' self-concept variables, predictabilities were nearly identical for the academic self-concept in physics (R^2 = .49) and the general academic self-concept (R^2 = .48).

Response consistency was high across all assessed constructs, ranging from r = .69 to r = .72, that is, perceivers applied their cue knowledge consistently and utilized the cues in a comparable manner across targets to judge targets' characteristics. In other words, perceivers used more or less the same linear decision rule for a specific criterion for each target, essentially giving the same weight to the same cue across all targets.

Perceivers' *cue sensitivity* values for the four constructs ranged from r = -.09 for intelligence to r = .37 for intrinsic motivation. This indicates that perceivers did not use valid cues when judging intelligence, but to some extent they used valid information in the case of intrinsic motivation and the academic self-concept in physics. In accordance with the fact that the highest accuracy outcomes were obtained for intrinsic motivation in physics, cue sensitivity scores were also highest for this construct, followed by the academic self-concept in physics.

Predictability, response consistency, and cue sensitivity are summary measures of the lens model that are all based on the availability and utilization of cues. In the following, we present results for cue utilization and cue validity that provide a more detailed insight. The results are based on multiple regressions with the cues as independent variables and either judgment or criterion of each of the four constructs as dependent variables. The regression analyses included both static and dynamic cues.

4.3 Cue Utilization and Cue Validity

Cue utilization and cue validity scores for each construct are displayed in Table 4.

4.3.1 Cue Utilization

Which cues were given particular importance when judging different target characteristics? In other words, what kind of information did perceivers rely on and how much weight did they give to different kinds of information?

Overall, the selected set of cues significantly predicted perceivers' judgments. Hence, perceivers relied on the behavioral and physical cues in their judgments of students' general (R^2 = .43) and subject-specific academic self-concept (R^2 = .58), intrinsic motivation in physics (R^2 = .55) and intelligence (R^2 = .44). In the case of single perceivers, cue utilization values for the individual cues were rather low. Regarding the cues describing a student's physical appearance (i.e., static cues), the cue utilizations of the cues sex, dark appearance, distinctive face, mature face, and eyeglasses were significantly different from zero for all four examined constructs. That is, perceivers rated targets' self-concept, motivation, and intelligence higher when the targets had a dark appearance and wore eyeglasses, and rated targets with a distinct and mature appearing face lower. A cue that contributed especially strong to perceivers' judgments of all four constructs was targets' sex: Male targets were perceived as having a higher academic self-concept, a higher intrinsic motivation, and a higher intelligence. For the case of students' behavior (i.e., dynamic cues), perceivers regarded students' attentive and self-assured, non-shy facial expression as well as expressive and tensed gestures as indicators for higher levels of all four assessed constructs. All these relations were positive and there is also no indication for a differential use of these cues across the four constructs. Average-perceiver cue utilizations are consistent with these results. On a descriptive level, relations for the average perceiver were stronger than the coefficients obtained from the single-perceiver regression analysis described above. Average-perceiver regression coefficients were particularly high and significant for students' attentive facial expressions across all constructs. For the academic self-concept in physics, a self-assured facial expression moreover significantly predicted the perceiver judgment. A dark appearance, as well as wearing eyeglasses and students' sex were significant predictors across all investigates constructs, except for wearing eyeglasses in relation to students' intelligence and a dark appearance in relation to students' intrinsic motivation in

Table 4

Cue Validity (CV and CV_{Corr}), Single-Perceiver (CU), and Average-Perceiver (CU_{Corr}) Cue Utilizations for Students' Broad Academic Self-Concept, Academic Self-Concept Physics, Intrinsic Motivation Physics, and Intelligence

	Cue validity (CV)	CV_{Corr}	Cue utilization CU)	CU_{Aver}	CU_{Corr}	Cue validity (CV)	CV_{Corr}	Cue Utilization CU)	CU_{Aver}	CU_{Corr}
Cue Measure	*A. Broad academic self-concept*					B. Academic Self-Concept Physics				
Attractiveness	-0.22	-.05	0.03	0.05	.12	-0.23	-.02	0.01	0.03	.05
Masculinity	0.02	-.02	-0.02	-0.04	-.01	0.17	.05	0.00	0.00	.05
Dark appearance	-0.12	-.07	**0.11**	**0.26**	.13	-0.02	-.16	**0.13**	**0.24**	.07
Distinctive cloth. style	-0.29	-.32	**0.04**	0.08	.01	**-0.37**	**-.35**	0.02	0.05	.02
Distinctive hair	0.00	-.09	**-0.07**	-0.16	-.24	**0.36**	.08	**-0.04**	-0.09	-.19
Distinctive face	**-0.45**	-.29	**-0.07**	-0.17	-.19	-0.22	-.26	**-0.02**	-0.22	-.19
Mature face	0.17	.01	**-0.05**	-0.12	-.17	-0.06	-.04	**-0.05**	-0.09	-.10
Body size	0.18	.07	0.02	0.06	.09	**0.34**	.11	**0.08**	0.13	.15
Apparent social backg.	-0.02	.10	0.01	0.02	.18	0.27	.15	-0.02	-0.04	.09
Eyeglasses	0.18	-.02	**0.10**	**0.25**	.22	0.06	-.05	**0.12**	**0.23**	.22
Sex (boys: 0, girls: 1)	**-0.34**	-.31	**-0.22**	**-0.50**	**-.50**	**-0.33**	**-.33**	**-0.34**	**-0.64**	**-.64**
Facial expression										
Friendly	**0.52**	.06	-0.03	-0.10	.11	-0.03	.06	0.02	-0.13	.07
Self-assured, non-shy	0.07	.02	**0.12**	0.29	**.36**	0.07	.01	**0.10**	**0.19**	.29
Communicative	-0.05	-.04	**0.01**	0.02	.10	-0.07	.00	0.03	0.05	.14
Attentive	-0.06	.02	**0.10**	**0.25**	**.31**	0.10	.04	**0.15**	**0.29**	**.33**
Gestures										
Expressive	-0.14	.03	**0.06**	0.16	**.30**	-0.16	-.10	**0.07**	0.15	**.29**
Tensed	**0.42**	.23	**0.05**	0.12	-.07	**0.27**	.24	**0.04**	0.07	-.08
	C. Intrinsic Motivation Physics					D. Intelligence				
Attractiveness	-0.19	-.13	-0.01	-0.01	.05	-0.38	-.19	-0.02	-0.03	.09
Masculinity	**0.32**	.25	0.02	0.03	.04	0.32	**.31**	**-0.05**	-0.09	-.10
Dark appearance	-0.16	**-.29**	**0.11**	0.19	.01	-0.03	-.13	**0.17**	**0.38**	.16
Distinctive cloth. style	-0.22	-.05	**0.05**	0.09	.09	-0.25	-.19	0.00	-0.01	-.08
Distinctive hair	**0.29**	.14	**-0.06**	-0.12	-.19	0.08	-.05	-0.02	-0.03	-.19
Distinctive face	0.00	.05	**-0.08**	-0.15	-.10	0.18	-.08	**-0.15**	**-0.34**	**-.36**
Mature face	0.00	.12	**-0.06**	-0.11	-.13	-0.12	.01	**-0.12**	**-0.29**	**-.29**
Body size	**0.29**	.27	**0.07**	0.14	.19	0.09	-.04	**0.06**	0.12	-.02
Apparent social backg.	0.10	-.07	0.00	-0.01	.12	0.20	-.06	0.00	0.00	.21
Eyeglasses	0.01	-.04	**0.12**	**0.22**	.23	-0.12	-.06	**0.06**	0.16	.14
Sex (boys: 0, girls:1)	**-0.48**	**-.31**	**-0.30**	**-0.55**	**-.55**	0.05	.05	**-0.18**	**-0.42**	**-.42**
Facial expression										
Friendly	0.07	.04	**-0.05**	0.01	.22	-0.16	-.23	-0.05	-0.09	.07
Self-assured, non-shy	0.00	-.02	**0.08**	0.16	.28	0.14	.05	**0.07**	0.17	.26
Communicative	-0.03	.12	0.00	-0.01	.17	-0.20	-.22	-0.02	-0.04	.12
Attentive	0.20	.27	**0.21**	**0.39**	**.47**	0.09	-.04	**0.10**	**0.37**	**.36**
Gestures										
Expressive	0.04	.23	**0.07**	0.14	**.39**	-0.14	-.04	**0.06**	0.11	.17
Tensed	0.25	.14	**0.07**	0.13	-.06	0.00	-.04	**0.05**	0.17	.03

Note. $N_{Perceiver}$ = 102; N_{Target} = 45. CV Cue Validity (regression coefficients), CV_{Corr} Correlations Cues-Criterion values, CU Single Perceiver Utilization (regression coefficients), CU_{Aver} Average Perceiver Cue Utilizations (regression coefficients), CU_{corr} Correlations Cues-Average Criterion values. Bold coefficients $p \leq .05$ (*t*-Test, two-tailed).

physics. For the case of intelligence, a distinctive and mature face played a significant role for the perceiver judgments, in addition to an attentive facial expression.

4.3.2 Cue Validity

Which cues were valid for students' academic self-concept, motivation and intelligence and which differences in cue validity could be identified across the four investigated student characteristics?

Targets' sex had a rather strong association with students' academic self-concept (both general and domain specific for physics) and intrinsic motivation in physics: Girls in our sample had lower self-reported self-concept and motivation compared to the boys. Targets' sex was unrelated to intelligence. Despite controlling for targets' sex, body size (i.e., being tall), as well as masculinity were cues with a substantial validity for intrinsic motivation in physics ($b \geq .30$). Controlling a characteristic for targets' sex means that boys and girls are only compared within their sex group irrespective of the difference in the group means. Hence, taller boys as well as taller girls and more masculine appearing boys as well as girls indicated higher intrinsic motivation levels. Body size was also predictive for the domain-specific self-concept in physics, alongside distinctive hair. Distinctive hair was referring to the hair style of a student and indicated to what extent the hair was rather styled in an unobtrusive or flamboyant manner. Beyond target's hair style, also the distinctiveness or peculiarity of a target's clothing style was a valid cue, however, it was significantly associated with a lower academic self-concept in physics. Among the dynamic cues, a friendly facial expression, as well as tensed gestures positively predicted students' values on the general academic self-concept scale.

Taken together, differences in judgment accuracy across the four investigated student characteristics could be associated with perceivers' utilization of more or less valid cues. Such cues could contribute positively or negatively the student characteristic. For students' intelligence, the construct with the lowest accuracy outcomes, there were no valid cues available. As a consequence, perceivers relied on other, non-relevant information. Positive weights were allocated to an attentive facial expression, a dark appearance, being tall, wearing eyeglasses, a self-assured facial expression and expressive and tensed gestures. Negative weights were given to sex, masculinity and a distinct and mature face. In sum, boys as well as students appearing masculine, average and juvenile within their sex group were judged as more intelligent.

For the case of students' broad academic self-concept, which was also judged with comparatively low accuracy, some valid cues were available to the perceivers. Such cues involved students' sex and distinctive face, which negatively predicted students' self-concept, whereas students' friendly expression and tensed gestures were associated with higher broad self-concept outcomes. However, apart from sex, perceivers did not allocate equal much weight to such valid cues, which may have led to the rather low accuracy. In particular, perceivers did not rely on students' friendliness as an important cue and also did not allocate sufficient importance to tensed gestures of the students in the video. Instead, perceivers relied on several other sources of information which were not valid, such as an attentive or self-assured facial expression or whether or not a student was wearing eyeglasses. For the case of the two subject related student characteristics, i.e., academic self-concept and intrinsic motivation in physics, the models of the perceivers were more closely related to the model of the environment, which also became visible in the predictability and cue sensitivity outcomes. In the case of students' academic self-concept in physics, perceivers were able to identify the importance of students' body size, sex and tensed gestures but missed out on the negative relationship of distinctive clothes and positive relationship of distinctive hair with students' academic self-concept in physics. For students' intrinsic motivation in physics, which was perceived with the highest accuracy, sex was the most valid cue with the largest negative effect size across all investigated characteristics, followed by masculinity, being tall and a distinctive hair style. Perceivers were able to identify the relationship of being male and tall with higher intrinsic motivation values in physics, but simultaneously misinterpreted other student characteristics as relevant indicators for students' intrinsic motivation in physics, in particular students' attentive facial expression, eyeglasses and a dark physical appearance.

5 Discussion

Based on BLM, we investigated the accuracy of teachers' judgments about students' self-concept, motivation, and intelligence in a zero-acquaintance situation. Student characteristics, such as the above-mentioned, are distal variables that cannot be observed directly but must be inferred by using proximal variables, that is, appearance-related or behavioral cues. In BLM, accurate judgments result if perceivers consistently utilize cues in their judgments that actually predict students' characteristics. The cues that were available in the brief videos shown to the perceivers,

are characteristics of body appearance, gestures, and facial expression, but no audible or speech information. We wanted to find out if and to what extent these cues actually predict the criteria and are utilized by the perceivers when a zero-acquaintance and thin-slices-of-behavior approach is applied. By utilizing brief videos of students that were previously unknown to the perceivers we intended to focus on the judgment process per se and independent as well as isolated of possible previous teacher-student interaction or information about the students.

5.1 Judgment Accuracy

As was demonstrated by zero-acquaintance research in personality psychology, judgments are often accurate, even in situations with minimal information (Ambady et al., 1995; Ambady & Rosenthal, 1992; Back & Nestler, 2016; Murphy et al., 2015). In the present study, however, judgment accuracy outcomes were comparatively low across all investigated constructs, with an exception for students' intrinsic motivation, which was judged with moderate accuracy. This result is comparable to accuracy outcomes of intrinsic motivation judgments in previous classroom studies (Spinath, 2005; Urhahne et al., 2011). For intrinsic motivation, valid cue information may be easily recognizable and highly visible so that even a short exposition is sufficient for a rather accurate judgment. Knowing the student for a longer period of time does not seem to provide much additional benefit to judgment accuracy (Praetorius et al., 2015). With respect to students' self-concept, the accuracy outcomes in our study were lower than in the study by Praetorius et al. (2015). This difference may be due to the fact that the perceivers in the study by Praetorius and colleagues could in addition utilize verbal information from the targets. For intelligence, the low judgment accuracy we obtained in our study also differs from previous classroom research (Machts et al., 2016). Low accuracy outcomes for perceiver judgments of students' intelligence and the related overall weak relation of the selected set of cues with targets' measured intelligence could be due to the fact that the video material was nonverbal. Our finding corresponds to the lens model study conducted by Reynolds & Gifford (2001), who in their controlled experiment showed that intelligence can be more accurately perceived from auditory (e.g., speech rate and number of words) than from visual cues.

5.2 Predictability, Response Consistency, and Cue Sensitivity

Predictability, response consistency, and cue sensitivity (matching) all contribute to accuracy. First of all, a predictable environment is a basic prerequisite for accurate judgments. In our study, we found some evidence for predictability of the selected cues, and hence, some of the cues that were rated based on the brief videos of the students were able to predict the different characteristics of the students quite well. Because the environment can't be controlled by the perceiver, response consistency and cue sensitivity are the only means by which a perceiver can influence his or her judgment accuracy. Notably, response consistency was rather high, indicating that perceivers used their judgment strategy in a consistent manner across different targets and characteristics. However, being consistent in one's judgment strategy only leads to an accurate judgment, when this strategy is characterized by a perceivers' ability to select and utilize the cues that are valid for a certain characteristic (i.e., cue sensitivity). Cue sensitivity is high when the judgment model (i.e., perceiver judgments) matches the environmental model (i.e., students' actual self-report and standardized test data). In the present study we found differences in cue sensitivity between the individual constructs. Hence, perceivers were more receptive to relevant information in their judgments of students' intrinsic motivation compared to students' intelligence. In the case of intelligence, the judgment model apparently did not match the environmental model sufficiently enough. Considering the rather low accuracy outcomes in the present research overall, one can conclude that cue sensitivity being comparatively low could be the main factor explaining this result.

5.3 Cue Utilization

Lens model analyses showed that perceivers' judgments of the four student characteristics can be predicted by the utilization of certain cues. Whereas the majority of cues was utilized by the perceivers for their judgments across all criteria, however with different weights, some cues were only utilized for one or two of the investigated characteristics. For instance, a friendly facial expression was only assumed to play a role for students' academic self-concept in physics (however, negatively) and a distinctive clothing style was only expected to be related to students' intrinsic motivation in physics and the broad academic self-concept. However, looking friendly was actually related to the general academic self-concept and therefore this is a cue that would have

been needed to be utilized for the judgment of this construct instead. Our study moreover demonstrated that some cues generally seem to be favored in the judgment process by the perceivers, such as a self-assured and attentive facial expression, as well as expressive and tensed gestures and whether or not a student is wearing eyeglasses. The fact that wearing eyeglasses was considered as an important indicator across all constructs is in line with our expectations given that previous personality research has recognized the connection between wearing eyeglasses and the perception of intelligence or related constructs in individuals (Leder et al., 2011).

Among the cues underlying our analyses, students' sex is of special interest. Perceivers utilized this cue for their judgments of all four student characteristics, and cue validity of sex was particularly high for students' self-concept and motivation, but low and not significant for students' intelligence. Thus, we found evidence for a gender bias with respect to perceiver judgments of students' intelligence in favor of the boys in our sample.

This finding corresponds to research on teacher expectations and evaluation demonstrating that female students often receive less favorable ratings from teachers and pre-service teachers in STEM fields (Bonefeld et al., 2020; Holder & Kessels, 2017; Keller, 2001). For self-concept and intrinsic motivation, too, boys were rated more favorably than girls. But the female students in our sample in fact reported lower self-concept and motivation levels. This finding can be interpreted considering the *kernel of truth* hypothesis, which is based on the idea that stereotypes about a certain group can contain elements that actually correspond to specific characteristics of this group (Jussim et al., 2009, 2015). Hence, by evaluating students' sex as a potentially important source of information for their judgment of self-concept and motivation, perceivers displayed a certain degree of accuracy. In a nutshell, we observed a gender effect on both sides of the lens (i.e., cue validity and cue utilization). When interpreting the results, one also needs to take into account that the video setting took place in a physics context, where we traditionally deal with a masculinity attribution (Keller, 2001). Hence, this cue was utilized in accordance to societal expectation and in line with the masculinity attribution of the STEM subjects in general.

5.4 Cue Validity

Overall, cue validity was rather low in this study, while in certain dimensions and for certain constructs, some cues in fact served as valid indicators for the investigated criteria. For instance, cues were less powerful to predict students' intelligence compared to the other student

characteristics, which then led to lower accuracy outcomes. Consequently, the same set of cues can be more or less suitable for different criteria of interest and therefore, predictability of a set of cues, i.e., the degree to which the selected set of cues represent targets' actual characteristics, has shown to vary largely depending on the nature of the investigated construct. Interestingly, cues that were relevant for constructs with a focus on the physics domain (both intrinsic motivation in physics and the domain-specific academic self-concept) were related to sex despite the standardization of the cues within female and male students in the sample. Characteristics that are linked to students' sex include students' body size and masculinity. Although boys are taller than girls on average, for instance, body size is related to some of the constructs within both sex groups. Intrinsic motivation was the construct that most evidently showed the importance of cues that are related to students' sex, as in addition to sex as cue itself, masculinity and being tall (i.e., body size) were highly valid indicators for students' intrinsic motivation in physics within the two sex categories.

The suitability of cues to aid the perceiver judgment may also depend on the nature of the cues, e.g., whether they refer to an individual's appearance or behavior and whether or not speech is included in the video or the information is non-verbal (Breil et al., 2021; Reynolds & Gifford, 2001). In our study, the proportion of valid cues was comparatively higher among the static cues than among the dynamic cues.

5.5 Limitations and Further Directions

First, participants in our study, both perceivers and targets were either enrolled in a German University program or attending a German school. Hence, generalizability of our findings to other countries or geographical regions is somewhat limited. Future studies might investigate whether or not such findings can be replicated in other countries or different cultural contexts.

Second, the perceivers in our study were teacher students and psychology undergraduate students but not experienced teachers. Topics like personality, interaction, and judgment are not only important for experienced teachers, but are also important issues in both fields of study. It is an open question if and to what degree teaching experience has an additional benefit for zero-acquaintance judgments in a classroom situation. Therefore, in future research it would be interesting to include experienced teachers, to identify the possible role of teaching experience for this kind of judgment.

Third, the present research took place in the context of a physics experiment. Physics as a school subject, as much as other STEM disciplines, has shown to be particularly vulnerable to judgmental biases. The gender bias revealed in the present research underlines the importance of establishing awareness programs and training, particularly for teachers in the STEM fields. Such activities could also help increasing the share of females that choose a career path in the natural sciences, given that career choices of young students are influenced by the experiences they make during their time in school (Leaper & Starr, 2019). We moreover believe that replicating our findings in other subject areas, such as language or math is an important task for future research.

Fourth, the cues that were identified in this study were not valid for intelligence. An important subsequent research step would therefore be to identify which cues, other than the ones included in the present research, could be suitable to identify an individual's cognitive abilities.

By looking at teacher judgments based on minimal information, we attempt to mimic a situation in which a teacher meets a student for the first time. This approach is very fruitful in identifying processes involved in teacher initial impressions that can have an impact on professional judgment and subsequent classroom interaction. Future research should take a deeper look into the translation of such findings into actual teacher behavior in the classroom to bridge the gap between evidence regarding teacher first impression and subsequent teacher-student interaction and judgments, grading and evaluations that can have an essential impact on students' future academic and life prospects.

6 Conclusion

The present study revealed the potential of BLM and applying a zero-acquaintance and thin-slice of behavior research design to shed light on the judgment process itself, which has received less attention in previous research on teacher judgment accuracy. We were thereby able to identify cues that are valid for students' various characteristics relevant for learning and performance in school and the degree to which perceivers are susceptible to such information in their judgment. Whereas judgment accuracy was overall rather low, lens model parameter analysis offered valuable insights showing that an accurate teacher judgment depends not only on the available cues, but also on the extent to which they are utilized by the perceivers. For constructs with low (intelligence) and higher (intrinsic motivation in physics) accuracy, these outcomes were reflected in lower sensitivity values and also a lower predictability of the applied set of cues.

Through looking at pre-service teachers' utilization of cues in their judgments, we were able to identify a gender bias, particularly for students' intelligence in favor of the male students in our sample. Those findings are a reminder that gender discrimination, which has traditionally been present in the STEM disciplines, is an ongoing issue that needs further attention in research and practice.

References

Ambady, N., Hallahan, M., & Rosenthal, R. (1995). On judging and being judged accurately in zero-acquaintance situations. *Journal of Personality and Social Psychology*, *69*(3), 518–529. https://doi.org/10.1037/0022-3514.69.3.518

Ambady, N., & Rosenthal, R. (1992). Thin slices of expressive behavior as predictors of interpersonal consequences: A meta-analysis. *Psychological Bulletin*, *111*(2), 256–274. https://doi.org/10.1037/0033-2909.111.2.256

Back, M. D., & Nestler, S. (2016). Accuracy of judging personality. In J. A. Hall, M. Schmid Mast, & T. V. West (Eds.), *The Social Psychology of Perceiving Others Accurately* (pp. 98–125). Cambridge University Press. https://doi.org/10.1017/CBO9781316181959.005

Back, M. D., Schmukle, S. C., & Egloff, B. (2011). A closer look at first sight: Social relations lens model analysis of personality and interpersonal attraction at zero acquaintance. *European Journal of Personality*, *25*(3), 225–238. https://doi.org/10.1002/per.790

Bergold, S., & Steinmayr, R. (2023). Teacher judgments predict developments in adolescents' school performance, motivation, and life satisfaction. *Journal of Educational Psychology*, *115*(4), 642–664. https://doi.org/10.1037/edu0000786

Bhowmik, C. V., Nestler, S., Schrader, F.-W., Praetorius, A.-K., Biesanz, J. C., & Back, M. D. (2021). Teacher judgments at zero-acquaintance: A social accuracy analysis. *Contemporary Educational Psychology*, *65*, Article 101965. https://doi.org/10.1016/j.cedpsych.2021.101965

Bonefeld, M., Dickhäuser, O., & Karst, K. (2020). Do preservice teachers' judgments and judgment accuracy depend on students' characteristics? The effect of gender and immigration background. *Social Psychology of Education*, *23*(1), 189–216. https://doi.org/10.1007/s11218-019-09533-2

Borkenau, P., Leising, D., & Fritz, U. (2014). Effects of communication between judges on consensus and accuracy in judgments of people's intelligence. *European Journal of Psychological Assessment*, *30*(4), 274–282. https://doi.org/10.1027/1015-5759/a000188

Breil, S. M., Osterholz, S., Nestler, S., & Back, M. D. (2021). Contributions of nonverbal cues to the accurate judgment of personality traits. In T. D. Letzring & J. S. Spain (Eds.), *The*

Oxford handbook of accurate personality judgment (pp. 195–218). Oxford University Press. https://doi.org/10.31234/osf.io/mn2je

Brunswik, E. (1956). *Perception and the Representative Design of Psychological Experiments*. University of California Press. https://doi.org/10.1525/9780520350519

Cooksey, R. W., Freebody, P., & Davidson, G. R. (1986). Teachers' Predictions of Children's Early Reading Achievement: An Application of Social Judgment Theory. *American Educational Research Journal*, *23*(1), 41–64. https://doi.org/10.3102/00028312023001041

Darley, J. M., & Fazio, R. H. (1980). Expectancy confirmation processes arising in the social interaction sequence. *American Psychologist*, *35*(10), 867–881. https://doi.org/10.1037/0003-066X.35.10.867

Fletcher, T. D. (2022). *QuantPsyc: Quantitative Psychology Tools* (1.6) [Computer software]. https://cran.r-project.org/web/packages/QuantPsyc/index.html

Förster, N., & Böhmer, I. (2017). Das Linsenmodell – Grundlagen und exemplarische Anwendungen in der pädagogisch-psychologischen Diagnostik [The lens model – fundamental aspects and exemplary applications in educational assessment]. In A. Südkamp & A.-K. Praetorius (Eds.), *Diagnostische Kompetenz von Lehrkräften. Theoretische und methodische Weiterentwicklungen [Diagnostic competence of teachers. Theoretical and methodological advancements]* (pp. 46–50). Waxmann.

Frey, A., Taskinen, P. H., Schütte, K., Prenzel, M., Artelt, C., Baumert, J., Blum, W., Hammann, M., Klieme, E., & Pekrun, R. (Eds.). (2009). *PISA 2006 Skalenhandbuch. Dokumentation der Erhebungsinstrumente [PISA 2006 scale documentation]*. Waxmann.

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

Funder, D. C., & Ozer, D. J. (2019). Evaluating Effect Size in Psychological Research: Sense and Nonsense: *Advances in Methods and Practices in Psychological Science*, *2*(2), 156–168. https://doi.org/10.1177/2515245919847202

Hardy, I., Decristan, J., & Klieme, E. (2019). Adaptive teaching in research on learning and instruction. *Journal for Educational Research Online*, *11*(2), 169–191.

Harris, M. J., & Garris, C. P. (2008). You never get a second chance to make a first impression: Behavioral consequences of first impressions. In N. Ambady & J. J. Skowronski (Eds.), *First impressions* (pp. 147–168). Guilford Publications.

Helmke, A., & Schrader, F.-W. (1987). Interactional effects of instructional quality and teacher judgement accuracy on achievement. *Teaching and Teacher Education*, *3*(2), 91–98.

Holder, K., & Kessels, U. (2017). Gender and ethnic stereotypes in student teachers' judgments: A new look from a shifting standards perspective. *Social Psychology of Education*, *20*(3), 471–490. https://doi.org/10.1007/s11218-017-9384-z

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*(2), 131-155. https://doi.org/10.1207/s15327957pspr0902_3

Jussim, L., Cain, T. R., Crawford, J. T., Harber, K., & Cohen, F. (2009). The unbearable accuracy of stereotypes. In T. D. Nelson (Ed.), *Handbook of prejudice, stereotyping, and discrimination* (pp. 199–227). Psychology Press.

Jussim, L., Crawford, J. T., & Rubinstein, R. S. (2015). Stereotype (In)Accuracy in Perceptions of Groups and Individuals. *Current Directions in Psychological Science*, *24*(6), 490–497. https://doi.org/10.1177/0963721415605257

Karelaia, N., & Hogarth, R. M. (2008). Determinants of linear judgment: A meta-analysis of lens model studies. *Psychological Bulletin*, *134*(3), 404–426. https://doi.org/10.1037/0033-2909.134.3.404

Keller, C. (2001). Effect of Teachers' Stereotyping on Students' Stereotyping of Mathematics as a Male Domain. *The Journal of Social Psychology*, *141*(2), 165–173. https://doi.org/10.1037/amp0000263

Kriegbaum, K., Jansen, M., & Spinath, B. (2015). Motivation: A predictor of PISA's mathematical competence beyond intelligence and prior test achievement. *Learning and Individual Differences*, *43*, 140–148. https://doi.org/10.1016/j.lindif.2015.08.026

Lansu, T. A. M., & van den Berg, Y. H. M. (2020). Thin-Slice Judgments of Children's Social Status and Behavior. *The Journal of Experimental Education*, *90*(4), 1–14. https://doi.org/10.1080/00220973.2020.1808943

Leaper, C. & Starr, C. R. (2019). Helping and hindering undergraduate women's STEM motivation: experiences with STEM encouragement, STEM-related gender bias, and

sexual harassment. *Psychology of Women Quarterly*, *43*(2), 165–183. https://doi.org/10.1177/0361684318806302

Leder, H., Forster, M., & Gerger, G. (2011). The glasses stereotype revisited: Effects of eyeglasses on perception, recognition, and impression of faces. *Swiss Journal of Psychology / Schweizerische Zeitschrift Für Psychologie / Revue Suisse de Psychologie*, *70*, 211–222. https://doi.org/10.1024/1421-0185/a000059

Liepmann, D., Beauducel, B., Brocke, B., & Nettelstroh, W. (2012). *IST-Screening. Intelligenz-Struktur-Test [IST-Screening. Intelligence-Structure-Test]*. Hogrefe.

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

Marksteiner, T., Reinhard, M.-A., Dickhäuser, O., & Sporer, S. L. (2012). How do teachers perceive cheating students? Beliefs about cues to deception and detection accuracy in the educational field. *European Journal of Psychology of Education*, *27*(3), 329–350. https://doi.org/10.1007/s10212-011-0074-5

Murphy, N. A. (2007). Appearing smart: The impression management of intelligence, person perception accuracy, and behavior in social interaction. *Personality and Social Psychology Bulletin*, *33*(3), 325–339. https://doi.org/10.1177/0146167206294871

Murphy, N. A., Hall, J. A., & Colvin, C. R. (2003). Accurate intelligence assessments in social interactions: Mediators and gender effects. *Journal of Personality*, *71*(3), 465–493. https://doi.org/10.1111/1467-6494.7103008

Murphy, N. A., Hall, J. A., Ruben, M. A., Frauendorfer, D., Schmid Mast, M., Johnson, K. E., & Nguyen, L. (2019). Predictive Validity of Thin-Slice Nonverbal Behavior from Social Interactions. *Personality & Social Psychology Bulletin*, *45*(7), 983–993. https://doi.org/10.1177/0146167218802834

Murphy, N. A., Hall, J. A., Schmid Mast, M., Ruben, M. A., Frauendorfer, D., Blanch-Hartigan, D., Roter, D. L., & Nguyen, L. (2015). Reliability and validity of nonverbal thin slices in social interactions. *Personality & Social Psychology Bulletin*, *41*(2), 199–213. https://doi.org/10.1177/0146167214559902

Möller, J., Pohlmann, B., Köller, O., & Marsh, H. W. (2016). A meta-analytic path analysis of the internal/external frame of reference model of academic achievement and academic

self-concept. *Review of Educational Research*, *79*(3), 1129–1167. https://doi.org/10.3102/0034654309337522

Nestler, S., & Back, M. D. (2013). Applications and Extensions of the Lens Model to Understand Interpersonal Judgments at Zero Acquaintance. *Current Directions in Psychological Science*, *22*(5), 374–379. https://doi.org/10.1177/0963721413486148

Nestler, S., Egloff, B., Küfner, A. C. P., & Back, M. D. (2012). An integrative lens model approach to bias and accuracy in human inferences: Hindsight effects and knowledge updating in personality judgments. *Journal of Personality and Social Psychology*, *103*(4), 689–717. https://doi.org/10.1037/a0029461

Nickerson, R. S. (1998). Confirmation bias: A ubiquitous phenomenon in many guises. *Review of General Psychology*, *2*(2), 175–220. https://doi.org/10.1037/1089-2680.2.2.175

Praetorius, A.-K., Drexler, K., Rösch, L., Christophel, E., Heyne, N., Scheunpflug, A., Zeinz, H., & Dresel, M. (2015). Judging students' self-concepts within 30s? Investigating judgement accuracy in a zero-acquaintance situation. *Learning and Individual Differences*, *37*, 231–236. https://doi.org/10.1016/j.lindif.2014.11.015

Praetorius, A.-K., Karst, K., Dickhäuser, O., & Lipowsky, F. (2011). Wie gut schätzen Lehrkräfte die Fähigkeitsselbstkonzepte ihrer Schüler ein? [How accurate are teachers in their judgments of pupils' self-concepts?] *Psychologie in Erziehung Und Unterricht*, *2*, 81–91. https://doi.org/10.2378/peu2010.art30d

R Development Core Team. (2020). *R: A language and environment for statistical computing* [Computer software]. R Foundation for Statistical Computing.

Revelle, W. (2023). *psych: Procedures for Psychological, Psychometric, and Personality Research* (2.3.6) [Computer software]. https://cran.r-project.org/web/packages/psych/index.html

Reynolds, D. J., & Gifford, R. (2001). The sounds and sights of intelligence: A lens model channel analysis. *Personality and Social Psychology Bulletin*, *27*(2), 187–200. https://doi.org/10.1177/0146167201272005

Rost, D. A., Sparfeldt, J. R., & Schilling, S. R. (2007). *DISK-GITTER mit SKSLF-8. Differentielles Schulisches Selbstkonzept-Gitter mit Skala zur Erfassung des Selbstkonzepts schulischer Leistungen und Fähigkeiten* [Differential self-concept scale to assess the academic self-concept]. Hogrefe.

Schnitzler, K., Holzberger, D., & Seidel, T. (2020). Connecting Judgment Process and Accuracy of Student Teachers: Differences in Observation and Student Engagement Cues to Assess Student Characteristics. *Frontiers in Education*, *5*, Article 602470. https://doi.org/10.3389/feduc.2020.602470

Seidel, T., Prenzel, M., Duit, R., & Lehrke, M. (2003). *Technischer Bericht zur Videostudie "Lehr-Lern-Prozesse im Physikunterricht"* [*Technical Report of the Video Study "Teaching and Learning Processes in the Physics Classroom"*] (1., Aufl.). IPN Leibniz-Institut f. d. Pädagogik d. Naturwissenschaften an d. Universität Kiel.

Spinath, B. (2005). Akkuratheit der Einschätzung von Schülermerkmalen durch Lehrer und das Konstrukt der diagnostischen Kompetenz [Accuracy of teachers' judgments regarding students' characteristics and the construct of diagnostic competence]. *Zeitschrift Für Pädagogische Psychologie*, *19*(1), 85–95. https://doi.org/10.1024/1010-0652.19.12.85

Stopfer, J. M., Egloff, B., Nestler, S., & Back, M. D. (2014). Personality Expression and Impression Formation in Online Social Networks: An Integrative Approach to Understanding the Processes of Accuracy, Impression Management and Meta–Accuracy. *European Journal of Personality*, *28*(1), 73–94. https://doi.org/10.1002/per.1935

Südkamp, A., Kaiser, J., & Möller, J. (2012). Accuracy of teachers' judgments of students' academic achievement: A meta-analysis. *Journal of Educational Psychology, 104*(3), 743–762. https://doi.org/10.1037/a0027627

Urhahne, D., Chao, S.-H., Florineth, M. L., Luttenberger, S., & Paechter, M. (2011). Academic self-concept, learning motivation, and test anxiety of the underestimated student. *The British Journal of Educational Psychology*, *81*(1), 161–177. https://doi.org/10.1348/000709910X504500

Wickham, H. (2022). *plyr: Tools for Splitting, Applying and Combining Data* (1.8.8) [Computer software]. https://cran.r-project.org/web/packages/plyr/index.html

CURRICULUM VITAE

Caroline V. Bhowmik
(geb. Wahle)

Wissenschaftliche Tätigkeiten

Wissenschaftliche Mitarbeiterin — Seit Oktober 2024

Department of Political, Historical, Religious and Cultural Studies
Universität Karlstad (KAU), Schweden

Gastwissenschaftlerin — September 2023 – April 2024

The Centre for Social Science Didactics (CSD)
Universität Karlstad (KAU), Schweden

Akademische Mitarbeiterin für Assessment — November 2015 – März 2016

Heidelberg School of Education – Universität Heidelberg

Wissenschaftliche Mitarbeiterin — März 2014 – September 2015

DFG - Graduiertenkolleg "Unterrichtsprozesse" (UPGRADE)
Rheinland-Pfälzische Technische Universität Kaiserslautern-Landau (RPTU)

Forschungsassistentin — April 2014 – Mai 2014

Self-Assessment Projekt (SAM) – RWTH Aachen

Studentische Hilfskraft — März 2008 – Oktober 2012

Institut für Geoinformatik (IFGI) – Universität Münster

Akademische Tutorin — Oktober 2011 – Januar 2012

Institut für Erziehungswissenschaft – Universität Münster

Akademische Tutorin und Studentische Hilfskraft — Oktober 2007 – Juli 2009

Institut für Erziehungswissenschaft – Universität Münster

Akademische Tutorin — November 2007 – Februar 2008

Institut für Geowissenschaften – Universität Münster

Auszeichnungen und Preise

ECP Best Poster Award — 2016
18th European Conference on Personality (ECP), Timisoara, Rumänien.

Promotionsstipendium — 2012
Deutsche Forschungsgemeinschaft (DFG)

Publikationen

Zeitschriftenartikel (Peer-Review)

Bhowmik, C.V., Nestler, S., Schrader, F.-W., Praetorius, A.-K., Biesanz, J., & Back, M.D. (2021). Teacher judgment accuracy at zero- acquaintance: A social accuracy analysis. *Contemporary Educational Psychology*, *65*, 101965 https://doi.org/10.1016/j.cedpsych.2021.101965

Bhowmik, C. V., Schrader, F., Back, M. D., & Steffen, S. (2021). An application of Brunswik's lens model to the educational context. *The Brunswik Society Newsletter*, 36, 18-20.

Santos P.B., Bhowmik C.V., Gurevych I. (2020). Avoiding Bias in Students' Intrinsic Motivation Detection. In: Kumar V., Troussas C. (eds) Intelligent Tutoring Systems. ITS 2020. Lecture Notes in Computer Science, vol 12149. Springer, Cham. https://doi.org/10.1007/978-3-030-49663-0_12

Qualifizierungsarbeit

Wahle, C. V. (2012). Brain Gain for Germany? A pilot study on Erasmus Mundus students' initial plans and actual choices. Unveröffentlichte Examensarbeit, Universität Münster.

Konferenzbeiträge

Symposien

Bhowmik, C.V., Nestler, S., Schrader, F.-W., Praetorius, A.-K., Biesanz, J., & Back, M.D. (2018, Juli). The role of teacher-student personality similarity for teacher judgment accuracy. Which traits are relevant? Beitrag im Rahmen des Symposiums: "Conceptual and measurement advances in the study of teachers' personality, social-emotional skills and effectiveness auf der 19th European Conference on Personality (ECP), Zadar, Kroatien.

Bhowmik, C.V., Nestler, S., Schrader, F.-W., Praetorius, A.-K., Biesanz, J., & Back, M.D. (2017, September). Welche Rolle spielen Ähnlichkeit, Sympathie und Attraktivität für das Lehrerurteil? Neue Ansätze in der Forschung zur diagnostischen Kompetenz am Beispiel des Social Accuracy Modells. Beitrag im Rahmen des Symposiums: "Lehrerurteile über Schülerinnen und Schüler: Aktuelle Befunde zu Einflussfaktoren und Konsequenzen" auf der Konferenz der Arbeitsgruppe "Pädagogische Psychologie (PAEPSY) der Deutschen Gesellschaft für Psychologie (DGPs), Münster.

Wahle, C.V., Back, M.D., Nestler, S., Schrader, F.-W., Pretsch, J., & Praetorius, A.-K. (2017, Juli). Understanding teacher judgments using linear models: an application of the lens model to teacher judgments of students' conscientiousness and motivation. Vortrag im Rahmen des Symposiums: "Shifting the focus from students to teachers: Individual differences in prospective and practicing teachers in high school and tertiary education" auf der 18th Conference of the International Society for the Study of Individual Differences (ISSID), Warschau, Polen.

Wahle, C.V., Back, M.D., Nestler, S., Pretsch, J., Schrader, F.-W., & Praetorius, A.-K. (2017, März). Welche Hinweisreize nutzen Lehrkräfte für Ihr Urteil und welche sollten sie nutzen? Eine Linsenmodellanalyse. Vortrag im Rahmen des Symposiums "Urteile über Schülerinnen und Schüler – Analysen zum Urteilsprozess und dessen Rahmenbedingungen" auf der 5. Tagung der Gesellschaft für Empirische Bildungsforschung (GEPF), Heidelberg.

EINGELADENE SYMPOSIEN

Wahle, C.V., Back, M.D., Nestler, S., Biesanz, J., Schrader, F.-W., Praetorius, A.-K., & Hochdörffer, K. (2016, Juli). Effects of liking and similarity on the accuracy of teachers' first impressions of students: A social accuracy analysis. Eingeladener Vortrag als Teil des Symposiums "Teacher Personality" auf der 18th European Conference on Personality (ECP). Timisoara, Rumänien.

INDIVIDUELLE VORTRÄGE

Bhowmik, C. V. (2024, January). Revised perspectives on Bildung in light of the Anthropocene in the context of Geography: A comparison between Sweden and Germany. Vortrag auf der second virtual online conference of the Education and Bildung in the Anthropocene network (EBAN), Karlstad, Schweden.

Wahle, C.V., Back, M.D., Nestler, S., Schrader, F.-W., Pretsch, J., Praetorius, A.-K. (2017, August). What drives teachers' judgments of students' characteristics? A lens model analysis. Vortrag auf der 17th Bienniel Conference of the European Association for Research on Learning and Instruction (EARLI). Tampere, Finnland.

Wahle, C.V., Back, M.D., Nestler, S., Schrader, F.-W., Praetorius, A.-K., & Biesanz, J. (2017, August). Perceiving students' characteristics accurately - what makes the good judge? Vortrag auf der 17th Bienniel Conference of the European Association for Research on Learning and Instruction (EARLI). Tampere, Finnland.

Wahle, C.V., Praetorius, A.K., Hochdörrfer, K., & Schrader, F.-W. (2015, August). Once a good judge, always a good judge? On the stability of pre-service teachers' judgment accuracy. Vortrag auf der 16th Bienniel Conference of the European Association for Research on Learning and Instruction (EARLI). Limassol, Zypern.

Hochdörffer, K., Wahle, C.V., & Schrader, F.-W. (2015, August). Actor and partner effects of student characteristics on learning outcomes in cooperative learning. Vortrag auf der 16th Bienniel Conference of the European Association for Research on Learning and Instruction (EARLI). Limassol, Zypern.

Wahle, C.V., Praetorius, A.-K., Hochdörffer, K., & Schrader, F.-W. (2015, July). Perceiving students' individual characteristics accurately based on minimal information: Effects of liking and teacher-student similarity. Panel discussant auf der International Society for the Study of Individual Differences (ISSID) conference. London, Kanada.

Sturm, N., Wahle, C.V., Rasch, R. & Schnotz, W. (2015, July). Self-generated representations are the key: The importance of external representations in predicting problem-solving success. Vortrag auf dem 39th meeting of the International Group for the Psychology of Mathematics Education (PME), Hobart, Australien.

Wahle, C.V., Hochdörffer, K. , Praetorius, A.-K., & Schrader, W.-F. (2014, September). Erfahrung macht den Meister? Untersuchungen zur Diagnosegenauigkeit von Lehramtsstudierenden. Vortrag auf der 79. Tagung der Arbeitsgruppe für Empirische Pädagogische Forschung (AEPF), Hamburg.

Weisenburger, K., Wahle, C. V. & Ludwig, P. (2013, September). Effects of (minimal) teacher interventions and learners' self-regulatory competence on learning processes and outcomes. Vortrag auf der Emerging Researchers' Conference of the European Conference on Educational Research (ECER), Istanbul, Türkei.

POSTERBEITRÄGE

Reinmöller, M., Wahle, C.V., & Schrader, F.-W. (2017, März). Diagnostische Kompetenz auf der Metaebene: Wie erklären sich Lehrkräfte ihr eigenes Urteil? Posterbeitrag auf der 5. Tagung der Gesellschaft für Empirische Bildungsforschung (GEPF), Heidelberg.

Wahle, C.V., Back, M.D., Nestler, S., Schrader, F.-W., Pretsch, J., Praetorius, A.-K., & Hochdörffer, K. (2016, July). Appearing smart, confident and motivated: A lens model approach to teachers' judgment accuracy. Mit dem "Best Poster Award" ausgezeichneter Posterbeitrag auf der 18th European Conference on Personality (ECP). Timisoara, Rumänien.

Hochdörffer, K., Wahle, C.V., & Schrader, F.-W. (2015, March). Gleich und Gleich gesellt sich gern? Effekte von homo- bzw. heterogener Zusammensetzung auf den Lernerfolg in dyadischen Schülerinteraktionen. Posterbeitrag auf der 3. Tagung der Gesellschaft für Empirische Bildungsforschung (GEPF), Bochum.

Wahle, C.V., Praetorius, A.-K., & Hochdörffer, K., & Schrader, F.-W. (2015, February). Are experienced teachers the better judges of students' characteristics based on thin slices of behavior? Posterbeitrag auf dem Annual Meeting of the Society for Personality and Social Psychology (SPSP), Long Beach, Kalifornien.

Hochdörffer, K. & Wahle, C.V., & Schrader, F.-W. (2015, February). Birds of a feather flock together? Effects of students' personality traits on academic achievement and dyadic interactions. Posterbeitrag auf dem Annual Meeting of the Society for Personality and Social Psychology (SPSP), Long Beach, Kaliforninen.

Wahle, C. V., Weisenburger, K., Weber, C. & Hellrigel, S. (2013, September). "Nicht für die Schule, für das Leben lernen wir!" - Lehrerinterventionen und Selbstregulationskompetenz von Schüler/-innen in selbstständigkeitsorientierten Lernarrangements. Posterbeitrag auf der 78. Konferenz der Arbeitsgruppe für Empirische Pädagogische Forschung (AEPF), Dortmund.

EINGELADENE VORTRÄGE

Bhowmik, C. V. (2023, November). Towards a better understanding of teacher judgment accuracy: An integrative approach. Vortrag im Rahmen der Seminarreihe des Forschungszentrums für die Didaktik der Sozialwissenschaften (Centre for Social Science Didactics (CSD)) an der Universität Karlstad, Schweden.

Wahle, C.V. (2015, Juli). Urteilsprozesse von Lehrkräften: Bisherige Forschungstätigkeit und mögliche Perspektiven. Vortrag im Kolloquium der Arbeitsgruppen Psychologie in Erziehung und Bildung und Sozialpsychologie der Universität Münster.

EIDESSTATTLICHE ERKLÄRUNG

Eidesstattliche Erklärung

Hiermit erkläre ich, Caroline Verena Bhowmik, geboren am 20.6.1985 in Neuss, dass ich die vorliegende Dissertation selbst angefertigt und alle dafür von mir benutzten Hilfsmittel in der Arbeit angegeben habe.

Darüber hinaus wurde die vorliegende Dissertation weder in ihrer Gänze, noch Teile hiervon als Prüfungsarbeit für eine staatliche oder andere wissenschaftliche Prüfung eingereicht. Auch wurden weder die gleiche oder eine andere Abhandlung zuvor bei einer anderen Hochschule als Dissertation eingereicht.

Karlstad, 28. Juni 2024

_____

Caroline Verena Bhowmik

CREDIT AUTHOR STATEMENTS

Manuscript: Bhowmik, C. V., Stang-Rabrig, J., Hellmann, K., & F. – W. Schrader (2021). Urteilsakkuratheit von Lehrkräften bei der Einschätzung nicht-kognitiver Variablen im Fach Physik: Geschlechtsbezogene Verzerrungseffekte.

CRediT Role	Author 1: [C. V. Bhowmik]	Author 2: [J. Stang]	Author 3: [K. Hellmann]	Author 4: [F.-W. Schrader]
1. Conceptualization	☒ --	☐ --	☐ --	☐ --
2. Data curation	☒ --	☐ --	☐ --	☐ --
3. Formal analysis	☒ lead	☒ supporting	☒ supporting	☐ --
4. Funding acquisition	☐ --	☐ --	☐ --	☐ --
5. Investigation	☒ --	☐ --	☐ --	☐ --
6. Methodology	☒ lead	☒ supporting	☒ supporting	☒ supporting
7. Project administration	☒ --	☐ --	☐ --	☐ --
8. Resources	☒ --	☐ --	☐ --	☐ --
9. Software	☒ --	☐ --	☐ --	☐ --
10. Supervision	☐ --	☐ --	☐ --	☒ --
11. Validation	☐ --	☐ --	☐ --	☒ --
12. Visualization	☒ equal	☒ equal	☒ equal	☐ --
13. Writing – original draft	☒ lead	☒ supporting	☒ supporting	☐ --
14. Writing – review & editing	☒ equal	☒ equal	☒ equal	☒ equal

Manuscript: Bhowmik, C. V., Nestler, S., Schrader, F. W., Praetorius, A.- K., Biesanz, J. C., & Back, M. D. (2021). Teacher judgments at zero-acquaintance: A social accuracy analysis. *Contemporary Educational Psychology*, *65*, 101965.

CRediT Role	Author 1: [C. V. Bhowmik]	Author 2: [S. Nestler]	Author 3: [F.-W. Schrader]	Author 4: [A. K. Praetorius]	Author 5: [J. C. Biesanz]	Author 6: [M. D. Back]
1. Conceptualization	☒ lead	☒ supporting	☒ supporting	☐ --	☐ --	☒ supporting
2. Data curation	☒ lead	☐ --	☐ --	☒ supporting	☐ --	☐ --
3. Formal analysis	☒ lead	☒ supporting	☒ supporting	☐ --	☒ supporting	☒ supporting
4. Funding acquisition	☐ --	☐ --	☐ --	☐ --	☐ --	☐ --
5. Investigation	☒ --	☐ --	☐ --	☐ --	☐ --	☐ --
6. Methodology	☒ lead	☒ supporting	☐ --	☐ --	☐ --	☐ --
7. Project administration	☒ lead	☐ --	☐ --	☒ supporting	☐ --	☐ --
8. Resources	☒ lead	☐ --	☐ --	☒ supporting	☐ --	☐ --
9. Software	☒ lead	☒ supporting	☐ --	☐ --	☐ --	☐ --
10. Supervision	☐ --	☐ --	☒ equal	☐ --	☐ --	☒ equal
11. Validation	☐ --	☒ equal	☒ equal	☐ --	☐ --	☒ equal
12. Visualization	☒ --	☐ --	☐ --	☐ --	☐ --	☐ --
13. Writing – original draft	☒ --	☐ --	☐ --	☐ --	☐ --	☒ supporting
14. Writing – review & editing	☒ lead	☒ supporting	☒ supporting	☐ --	☐ --	☒ supporting

Manuscript: Bhowmik, C. V., Back, M. D, Nestler, S., & Schrader, F. W. Appearing smart, confident and motivated: A lens model approach to teacher judgment accuracy.

CRediT Role	Author 1: [C. Bhowmik]	Author 2: [M. D. Back]	Author 3: [S. Nestler]	Author 4: [F.-W. Schrader]
1. Conceptualization	☒ --	☐ --	☐ --	☐ --
2. Data curation	☒ --	☐ --	☐ --	☐ --
3. Formal analysis	☒ lead	☒ supporting	☒ supporting	☒ supporting
4. Funding acquisition	☐ --	☐ --	☐ --	☐ --
5. Investigation	☒ --	☐ --	☐ --	☐ --
6. Methodology	☒ equal	☒ equal	☒ equal	☒ equal
7. Project administration	☒ --	☐ --	☐ --	☐ --
8. Resources	☒ --	☐ --	☐ --	☐ --
9. Software	☒ lead	☐ --	☒ supporting	☐ --
10. Supervision	☐ --	☒ equal	☐ --	☒ equal
11. Validation	☐ --	☒ equal	☒ equal	☒ equal
12. Visualization	☒ lead	☐ --	☐ --	☒ supporting
13. Writing – original draft	☒ lead	☐ --	☐ --	☒ supporting
14. Writing – review & editing	☒ lead	☒ supporting	☒ supporting	☒ supporting