Soc Psychol Educ (2013) 16:215–239
DOI 10.1007/s11218-013-9215-9

The impact of accountability on teachers' assessments of student performance: a social cognitive analysis

Sabine Krolak-Schwerdt · Matthias Böhmer · Cornelia Gräsel

Received: 17 July 2012 / Accepted: 13 February 2013 / Published online: 4 April 2013

Abstract Research on teachers' judgments of student performance has demonstrated that educational assessments may be biased or may more correctly take the achievements of students into account depending on teachers' motivations while making the judgment. Building on research on social judgment formation the present investigation examined whether the accountability of teachers has an influence on judgment formation. We predicted that unaccountable teachers would activate social categories and use them for the assessment, whereas accountable teachers' attention would be directed to individual attributes of students. Using secondary school teachers as participants, three studies investigating teachers' assessments, inferences and memory for students' attributes supported these hypotheses. Thus, accountability appears to be a moderator of social information processing and judgment formation in the domain of educational assessments.

Keywords Social cognition · Accountability · Student assessment

1 Introduction

A central aspect of teachers' professional competence is the ability to assess students' achievements adequately. Giving grades and marks is the prototypical task in this

This research was funded by the German Research Foundation (Deutsche Forschungsgemeinschaft, DFG; Grants KR 2162/3-1 und GR 1863/4-1).

S. Krolak-Schwerdt (✉) · M. Böhmer
Faculty of Humanities, Arts and Educational Sciences, University of Luxembourg, B.P. 2, 7201 Walferdange, Luxembourg
e-mail: sabine.krolak@uni.lu

C. Gräsel
Bergische Universität Wuppertal, Wuppertal, Gemany

context. Some of the teachers' tasks are similar to those of psychological assessment in that an explicit assessment has to be given. Besides giving grades, assessments for school placements or tracking decisions belong to these tasks. Other assessments are more implicit in that no specific assessment is required, but students' achievements are estimated intuitively. Examples are decisions made during class such as "calling on a particular student".

These assessments have substantial relevance for individual students and consequently, high competence in assessing students correctly is seen as a key skill for teachers and future teachers (Shepard 2006). But at the same time, a number of studies have shown that teachers' assessments of student performance frequently do not meet the criteria of measurement theory such as reliability and validity, but seem to be rather subjective (Givvin et al. 2001; Swanson 1986). Within educational systems where assessments are used to make decisions about a student's future academic career, this may contribute to problems of social segregation and may be harmful to the personal and later professional development of students (Alpert and Bechar 2008). The Third International Mathematics and Science Study (TIMSS; BMBF 2001) revealed deficiencies in teachers' assessments in mathematics and physics. In grading written exams, teachers use the class as a social reference system. In classes of students with high achievement levels, a particular student has to perform better to receive the same grade that he or she would receive in a class of students with lower academic achievement levels which also affects students' academic self-concepts (Marsh 1987; Marsh et al. 2000). Similarly, in their meta-analysis on teacher judgment accuracy Hoge and Coladarci (1989) as well as Südkamp et al. (2012) come to the conclusion that on the one hand teachers' judgment accuracy of students' performance is fairly high, but on the other hand teachers' judgments leave 57 up to 72 % of the variation of students' test performance unexplained "which leaves plenty of room for improvements" (Südkamp et al. 2012, p. 13).

According to research on teacher expectations, teachers may develop stereotypical expectations about students' achievements on the basis of socioeconomic or ethnic background or gender (Andrews et al. 1997; Brophy and Good 1974; Parks and Kennedy 2007; Pigott and Cowen 2000; Reyna 2000; Weiner 2000). Expectations of student performance may in turn lead teachers to behave differently toward different students (Babad 1993; Graham 1990). Teachers communicate their beliefs about students' abilities by their interpersonal behavior—for instance, through the emotions they convey (Weiner 2000). In some cases, then, teachers' expectations and attributions may lead to a self-fulfilling prophecy (Rosenthal and Jacobson 1968). Self-fulfilling prophecies in the classroom are small in general, but they are stronger among students from stigmatized social groups (Jussim and Harber 2005).

More recent research has yielded heterogeneous results and has emphasized the importance of moderator variables such as teachers' goals, motivations, and accountability. Biases in judgments due to expectations are more likely to occur when there is an incentive to confirm an expectation or a striving to rapidly reach a particular conclusion. Judgment biases are less likely when there is motivation to develop an accurate impression of the target person or when the perceiver's outcomes depend on the target person (see Jussim et al. 1996, for a review). For example, teachers' assessments of students' achievements become less biased when teachers have the goal of

improving students' achievements (Goldenberg 1992) or when assertive parents offer evidence that conflicts with teachers' expectations (Good and Nichols 2001). Furthermore, teachers' attention and memory may be either biased or relatively unaffected by expectations depending on their assessment goals. Van Ophuysen (2006) asked elementary school teachers to provide a placement decision for a number of students. When teachers were subsequently given inconsistent information, that is, information that went counter to their decisions, they were able to adapt their recommendations to the new information.

In sum, these findings suggest that teachers' goals may influence their processing of students' achievements and thus give rise to more accurate or biased judgments. However, the educational sciences have not developed theoretical explanations for the varying quality of teachers' assessments that have been investigated empirically. Most importantly, such explanations would be a necessary prerequisite for improving teachers' assessment competencies and, thus, for solving social problems of discrimination of students due to biased educational assessments.

On the other hand, social cognition research has developed elaborated theoretical assumptions on the question of how the processing of person attributes such as behaviors, beliefs, aptitudes, and so forth, may be affected by perceivers' goals and motivations in everyday impression and judgment formation. Thus, principles of judgment formation documented in this line of research may provide a useful theoretical framework for investigating teachers' judgment processes and for explaining the underlying mechanisms of when and why more accurate or biased assessments occur.

In applying social judgment research to the analysis of teachers' assessments, an assessment is conceptualized as a social judgment, which is the result of a cognitive process. Two modes of social information processing that can be differentiated from each other are widely discussed (Brewer et al. 1988; Chaiken and Ledgerwood 2011; Chen and Chaiken 1999; Fiske 2011; Fiske et al. 1999; Fiske and Neuberg 1990).

1.1 Category-based or heuristic processing

In this mode, the processing of person attributes is directed by activated social categories or stereotypes (Fiske 1993; Fiske and Neuberg 1990). In this case, the information that is given about a person is interpreted and encoded in terms of the activated category. Social categories may be activated by easily accessible features such as age, race, gender, and so forth, or they may be primed by previous information. Within the domain of teachers' assessments, the socio-economic background and immigration status of students are categories that teachers use frequently (Ormrod 2006). Furthermore, it has been demonstrated that teachers develop more specific student categories such as "No. 1 in class" or "hard-working student" during their numerous encounters with students in the classroom (Hofer 1981).

Once a category is activated, a large number of person attributes, which are represented in the category, become available. Thus, person information is encoded on the basis of the activated stereotype such that stereotype-consistent person attributes are accessible from memory and used for making a judgment. Category-based processing

involves little attention to the attributes of an individual person and thus requires minimal cognitive effort.

In category-based processing, person memory, judgment formation, and inferences are influenced by the activated category. The storage of the person information is based on the category. Judgment formation is guided by the category such that the target person is judged depending on the content of the category. Trait inferences about the person seem especially likely with this type of processing, and they appear very early when information about the person is encoded (Fiske and Taylor 2008).

Because category-based processing is relatively effortless and proceeds rapidly, it may be considered an efficient mode of processing (Bodenhausen et al. 1999; Fiske 1993). However, because this type of processing involves paying little attention to the attributes of the individual person, it is susceptible to biased judgments.

1.2 Attribute-based processing

In this mode, the given information about the person is processed more systematically with comparatively high cognitive effort. In contrast to category-based processing, the attributes of the individual target person are the focus of attention. Thus, attribute-based processing is characterized by more thoroughly taking into account the person's given attributes (Fiske and Neuberg 1990). Person memory may then more accurately reflect the given attributes. As compared to category-based processing, it may be further assumed that the judgment is formed by more systematically integrating the person's given attributes (Fiske and Neuberg 1990).

Because attribute-based processing is relatively time consuming and associated with cognitive effort, it is only used as a last resort, giving priority to category-based processing (Fiske 1993).

Empirical findings have provided evidence that these two modes of judgment formation can be differentiated from each other. Investigations that have analyzed information processing using the "Think Aloud" method have shown that participants collected a greater number of information pieces in the attribute-based processing mode as compared to the category-based mode. In addition, they produced more complex and differentiated characterizations of the target person (Lerner and Tetlock 1994). Thus, the attribute-based processing mode is associated with integratively more complex thought, and the corresponding judgment is more precise because it more correctly reflects the given information.

Dual process theories of social cognition provide a framework for integrating the two types of judgment formation. Central to these theories is the assumption that perceivers can shift from category-based to more complex cognitive processing in response to certain demands. Pivotal to our research question are the continuum model (Fiske et al. 1999; Fiske and Neuberg 1990) and the accountability theory (Lerner and Tetlock 2003). In the continuum model, the motivation of the perceiver is a fundamental moderator of judgment formation. High versus low motivation of the perceiver is induced by his/her accountability for the judgment or decision.

Starting with the assumption that people prefer least-effort solutions, which would give priority to category-based processing, the continuum model specifies the

following conditions of category-based thinking: (a) when the person information is consistent and a social category is available and (b) when the perceiver is relatively unmotivated to pay a large amount of attention to the target person's attributes. By contrast, information processing is attribute-based when (a) person attributes are inconsistent and difficult to comprehend and/or (b) the perceiver is highly motivated to attend to the target person's attributes. Increased motivation may result from being accountable for a judgment that has serious consequences for the target person or from pressure to justify the judgment to others or from high internal judgment standards set by the judge himself (Fiske and Neuberg 1990; Tetlock and Lerner 1999). These conditions induce the striving for a more accurate judgment, which takes into account as many information pieces as possible.

In sum, the two models posit how motivational and cognitive determinants interact in order to shape individual judgment and choice. Studies have documented that people who are highly motivated preferentially use the attribute-based strategy, whereas low motivation more likely induces the category-based mode (see e.g., Fiske 1993; Fiske and Neuberg 1990; Gollwitzer and Moskowitz 1996; Krolak-Schwerdt et al. 2009). Thus, there exists ample evidence that motivation and accountability, respectively, are moderators of information processing. Within this theoretical framework, it has also been shown that consistency of person information and availability of a social category affect the type of judgment formation: Consistent, stereotypical person attributes are more likely to induce category-based thinking than inconsistent, nonstereotypical cues (Fiske and Neuberg 1990).

The continuum model postulates some intermediate processing stages, which may be relatively more category-driven or attribute-oriented. Most notably, the model assumes a flexible use of both types of judgment formation: People are "motivated tacticians" who are able to select from different processing strategies in response to the actual situational demands (Fiske 1993). Which mode of processing is selected in a particular situation thus depends on (a) the consistency/typicality of the target person's attributes and (b) the accountability of the perceiver for the judgment at hand.

In the present research, we used these models from social cognition research to investigate whether teachers' processing of students' achievements and behaviors and their resulting assessments were differentially influenced by their accountability. It was predicted that unaccountable teachers would activate social categories and use these categories for the assessment, whereas the attention of accountable teachers would be directed to individual attributes of students. In Study 1, we examined whether accountability influences teachers' assessments of students' achievements and teachers' inferences on the traits and aptitudes of students. This allowed us to apply our knowledge about the determinants of category-based versus attribute-based judgment formation to the domain of educational achievement judgments. Study 2 examined whether accountability would have an influence on the moment when social categories are activated in the processing of information. Study 3 explored whether accountability would affect teachers' memory for students' attributes, which in turn might moderate findings on achievement assessments.

To summarize, the present studies tested four hypotheses:

Hypothesis 1 Teachers' achievement judgments of student performance will be biased by available social categories when they are not accountable for their judgment. Accountable teachers will produce assessments that more accurately take into account the person's given information irrespective of the availability of a category.

Hypothesis 2 Unaccountable teachers will generate inferences about students on the basis of the available category. For accountable teachers, no such inferences are expected.

Hypothesis 3 Unaccountable teachers will activate social categories about students during their encoding of the person information. With accountability, no category activation will occur during the encoding process.

Hypothesis 4 Unaccountable teachers' memory of students' attributes will be biased by an available category. Accountable teachers' memory will more accurately reflect the information given about a student.

As to the differentiation of Hypotheses 1 and 2 the term judgment always refers to an achievement judgment or assessment of student performance as a result of inferences in the following, whereas inferences may also comprise other domains such as social behavior or personality.

The major objective of the present research was to demonstrate that the accountability of teachers induces more category-based or more accurate assessments of students' achievements. In addition, the present research aimed to shed light on the cognitive processes that may underlie the variations in the quality of assessments in an applied setting. Finally, the dual process theories underlying the present research were focused mainly on the effect of perceivers' motivation in everyday impression and judgment formation, for example, in judging agreeableness, likability, and so forth, of other persons. Our research may provide some insight into the question of whether these theories also provide a solid base for explaining more formal professional judgment formation as compared to informal judgments.

2 Study 1

To investigate Hypotheses 1 and 2, we presented teachers case reports about students, which were displayed on a computer screen. Half of the participants were made accountable and the other half were made unaccountable through instructions. Availability of a social category was varied by informing half of the participants that the student had an immigration background, whereas the other half received no such information. We predicted that biased judgments (Hypothesis 1) and category-based inferences about students (Hypothesis 2) would occur when teachers were unaccountable, whereas accountable participants' judgments were expected to be unaffected by social categories.

2.1 Methods

2.1.1 Participants and design

Thirty-two German secondary school teachers (19 male, 13 female) participated in the study. On average, the participants were 48.9 years old (SD = 6.9) and had 22.7 years (SD = 7.9) of experience in their profession. They were randomly assigned to the conditions of a 2 (accountability: accountable vs. unaccountable) × 2 (social category: available vs. not available) between-participants factorial design.

2.1.2 Materials and procedure

Participants were run one at a time. Upon entering the laboratory, participants in the *accountable* condition read the following set of instructions:

> In this study, we are interested in how you make school track recommendations[1] for students. Please imagine the following scenario: You are the homeroom teacher of a student whose parents ask you for a parent–teacher meeting. The family is planning to move, which means that the child will have to change schools. The child's parents ask you to recommend the secondary school track that best reflects the child's current development. In the following, you will receive information about the student's behavior, as it has been observed by yourself and other teachers in class and during recess. Additionally, you will receive the student's exam grades in mathematics and German. Your task is to use this information in order to make a secondary school track recommendation (see footnote 1) for this student.

Participants in the *unaccountable* condition received the following set of instructions:

> In this study, we are interested in how you form your first impression of a student. Please imagine the following scenario: You are a homeroom teacher and a new student joins your homeroom class. In the following, you will receive information about the student's behavior, as it has been observed by yourself and other teachers in class and during recess. Additionally, you will receive the student's exam grades in mathematics and German. Your task is to use this information in order to form a first impression of this student and to judge his or her behavior and achievement.

Following the manipulation of accountability, participants were handed the case materials, which consisted of the behavior description of a student, a German language test, and a mathematics test of the student. The behavior description comprised 17 statements of interpersonal behaviors of the student in the classroom (e.g., "Max enjoys planning games for his classmates"). The description was constructed by using practical guidelines for formulating school reports (Langer et al. 1993). The German

[1] Note that Germany practices academic tracking in which students are divided among different types of schools that differ in terms of academic demands and future educational possibilities (e.g., university access).

language class test and the mathematics class test were constructed according to German curricula for sixth graders. The subjects German language and mathematics were selected because they are included in the studies of international student assessment (e.g., PISA; Baumert et al. 2001, 2002) and have outstanding importance for educational assessments in schools. The German class test consisted of a task for text comprehension, a multiple-choice grammar exercise, and an essay. The mathematics class test comprised 12 tasks on the topics of adding numbers, fractions, and converting scale units. The tasks of both tests and the task solutions of a fictitious student were constructed in such a way that both test solutions corresponded to a grade of 3.5 on a 7-point rating scale ranging from 1 (*very good*) to 7 (*very bad*).

To vary the availability of a social category, we presented information about the immigration status of the student before participants received the case materials described above. In an experimental condition, the fictitious student was given a Turkish first name, and participants were informed that his father is Turkish and was born in a small village. In a control condition, the student was given a German first name.

After reading the case materials, participants had to fill out a questionnaire on their biographical data as a distractor task. After completing the distractor task, participants had to summarize the case report about the student in their own words. These are instructions that are frequently used in the domain of text comprehension research and are especially suited for the analysis of inferences from person descriptions (Gernsbacher 1996; Wintermantel and Christmann 1983). The participants' final task was to assess the German language performance and the mathematics performance of the student on a 7-point rating scale ranging from 1 (*very good*) to 7 (*very bad*).

2.1.3 Prestudies

In order to guarantee the curricular and external validity of the materials and instructions, we conducted four prestudies.

Prestudy 1 investigated whether the instructions correspond to professional tasks of teachers in the school setting and differ in accountability using 18 secondary school teachers with at least 10 years of professional school experience. Participants received both sets of instructions. They had to decide whether the instructions were associated with accountability with a "yes," "partly," or "no" response. Eighteen participants identified the first set of instructions, which involved the school track recommendation for the student, with the task of counseling parents and making track decisions, as associated with high accountability. Seventeen participants identified the second set of instructions, which involved the formation of a first impression of the student, as reflecting informal assessments during the class and as associated with low accountability. Thus, the instructions corresponded with the intended difference in accountability.

Prestudy 2 served two purposes. First, it was designed as a pilot study to investigate our principal prediction that unaccountable teachers would use an available category for the performance assessment, whereas unaccountable teachers would use the individual attributes of the student. Second, Prestudy 2 investigated if the instructions really evoke different levels of accountability as a manipulation check. Thirty-six school teachers with at least 10 years of professional school experience participated

Table 1 Regression weights of the predictors performance tests, immigration status and interpersonal behavior to predict performance assessments as criterion (Prestudy 2)

Predictors	Accountability	
	Unaccountable	Accountable
Performance tests	2.85*	2.76*
Immigration status	0.57*	0.22
Interpersonal behavior	0.46	0.68*

Significant regression weights are indicated by an asterisk (*), $p < .05$

in the prestudy. They were randomly assigned to one of the instructions. After reading the instruction they were handed a set of 16 vignettes with student descriptions, each of which contained as cues the performance scores in mathematics and German as well as a variable representing interpersonal behavior. Half of the vignettes were presented with an immigration status, whereas the other half were presented with a German first name. Thus, the profile of each student description consisted of four cues which were test scores in mathematics and in German, interpersonal behavior and immigration status. Moreover, cues in the student descriptions were varied within participants. After reading each student description, participants had to rate the academic achievement of the student. To check our accountability manipulation, teachers were additionally asked for the perceived accountability during their work on the student descriptions on a percentage scale ranging from 0 to 100 %.

To investigate our principal prediction, we computed a multiple regression analysis for each accountability condition. Predictors were a compound performance test which was set up by the German language test and the mathematics test, the immigration status and the interpersonal behavior. The criterion was the academic achievement assessment. Table 1 shows the results. In both accountability conditions, teachers took the performance-related information into account when making their assessments. However, only unaccountable teachers' assessments were influenced by the immigration status, whereas accountable teachers took the interpersonal behavior as additional individual information into account. These results may be interpreted as a first indication that accountability might be a moderator of teachers' performance assessments.

To check whether our instructions induced different degrees of accountability, we compared the scale values of the two groups by a t test. In the accountability condition ($M = 88.67$) a higher degree of accountability was perceived than in the unaccountable condition ($M = 55.35$), $t\,(33) = 2.41, d = 0.81, p < .05$. Thus, the instructions indeed evoked different degrees of accountability. Teachers who were made unaccountable through instruction rated their perceived accountability during their work on the student descriptions in the middle range of the scale, whereas teachers who were made accountable judged their perceived accountability as very high.

Prestudy 3 had the aim of investigating whether the behavior description of the main study was neutral with respect to category activation. Ten secondary school teachers with at least 10 years of professional school experience were instructed to estimate whether the behavior description allowed for inferences about the social and

immigration status of the student and, if so, to specify the corresponding category. All participants answered that the materials did not imply such inferences and that they were not able to specify any category. Thus, the behavior description was not confounded with the activation of social categories and, specifically, did not interfere with the activation of immigration status.

Prestudy 4 was conducted to construct the German language and mathematics tests. Two experts of pedagogical content knowledge (cf. Shulman 1986) in the subjects of mathematics and the German language selected tasks from curricula and arranged the selected tasks into a mathematics test and a German language test, respectively. Eighty-seven German sixth graders worked on the mathematics test and 80 sixth graders on the German language test. From the resulting pool of task solutions, the ones that comprised minor mistakes were selected. The selected task solutions were arranged into the final mathematics test and German language test of the fictitious student. Twenty-four secondary school teachers and 23 senior students of educational sciences gave grades for the two tests on a scale ranging from 1 (*very good*) to 7 (*very bad*). On average, both fictitious tests were given a grade of 3.5. Consequently, the tests were in the middle range of performance and may thus potentially be shifted in grading to the more extreme performance range due to category.

2.1.4 Dependent measures and coding

In order to analyze participants' summaries of the student descriptions, it was first necessary to distinguish between (a) phrasing that had a direct linguistic link to the initially presented information in that it completely or partially repeated this information, and (b) phrasing that went beyond the presented information in the form of inferences. Only this second type of phrasing was of interest within the summarization task and was thus considered in the analyses. Summaries were divided into propositional entities. Those propositions that showed no direct link to the original descriptions in the sense of being repetitions of or synonyms for the original text were coded as inferences. The second step of the analysis involved differentiating between types of inferences. This was done using the coding system by Wintermantel and Krolak-Schwerdt (2002), which differentiates between behavioral, personality, and achievement inferences on the basis of the attribution theory approaches of Kelley (1967) and Gilbert (1998). These categories allowed for a complete classification of all inferences. Two independent coders blind to the experimental hypotheses coded the summarization protocols. The agreement between coders (Cohen's Kappa) was $\kappa = .98$. When coders disagreed about a categorization, they discussed the item until reaching a consensus. The third step of analysis consisted of identifying whether each inference fit the activated category or not. Neutral inferences were coded the same as category-inconsistent inferences.

2.2 Results and discussion

To address the predictions regarding teachers' assessments of the student performance, a 2 (accountability) $\times$ 2 (category availability) MANOVA was conducted on the

Table 2 Ratings of German language performance and mathematics performance as a function of accountability and category availability (Study 1)

Accountability	Category availability	
	Available	Not available
German language performance		
Accountable	3.12	3.00
Unaccountable	4.00	2.83
Mathematics performance		
Accountable	3.38	3.25
Unaccountable	3.50	2.50

German language performance and mathematics performance were assessed on 7-point rating scales ranging from 1 (*very good*) to 7 (*very bad*)

German language performance ratings and the mathematics performance ratings. The interaction of accountability and category availability was significant, Wilks $\lambda = .79$, $\eta^2 = .21$, $p = .04$. No other effects were significant. As can be seen in Table 2, the pattern of means was consistent with the predictions. Planned contrasts revealed that participants in the unaccountable/category available condition were biased by the social category when assessing German language performance as compared to the unaccountable/no category available condition, $t(14) = -3.21$, $d = 1.61$, $p = .006$. The same held true for the assessments of mathematics performance, $t(14) = -2.33$, $d = 1.17$, $p = .04$. For accountable participants, there was no significant difference between the category available versus the not available condition, $t(14) = -0.23$, *ns*, for German language performance or for mathematics performance, $t(14) = 1.67$, *ns*.

Comparing the performance judgments of the teachers with the adequate assessments that were 3.5 for both the German language test and the mathematics test due to the construction of the materials demonstrated that unaccountable teachers' judgments deviated from the correct assessments to a larger extent than accountable teachers' judgments, and this held true across experimental conditions.

A 2 (accountability) $\times$ 2 (category availability) ANOVA was then conducted on the total number of inferences in the summarization protocols. The interaction of accountability and category availability was not significant, $F < 1$. However, there was a significant main effect of accountability, $F(1, 28) = 10.60$, $\eta^2 = .28$, $p = .003$, indicating that participants generated more inferences in the unaccountable condition ($M = 6.25$) as compared to the accountable condition ($M = 2.19$). Furthermore, there was a significant main effect of category availability, $F(1, 28) = 7.05$, $\eta^2 = .20$, $p = .01$, where more inferences were generated in the category available condition ($M = 5.87$) than in the no category available condition ($M = 2.56$).

Finally, an ANOVA with accountability as a single factor was conducted on the number of category-consistent inferences for the category available condition only. This revealed a main effect of accountability, $F(1, 14) = 6.64$, $\eta^2 = .32$, $p = .02$, which was consistent with the predictions. Participants who were not accountable for their judgments produced a much higher number of inferences that were consistent with the available category ($M = 4.25$) than accountable participants ($M = 0.56$).

A consistent pattern of results emerged from the major dependent variables: assessments of German language and mathematics performance, total number of inferences, and number of category-consistent inferences. Unaccountable participants assessed the performance of the student less favorably when the student was presented as Turkish and thus had an immigration status as compared to the assessments of the same case materials pertaining to a German student without an immigration background. Furthermore, unaccountable participants drew a considerable number of inferences about the student that went beyond the given information irrespective of having a social category at hand. Thus, even when there was no social category available, unaccountable participants generated attributions of the student's personality, behavior, and aptitude. An additional analysis of the number of inferences in the category available condition revealed that unaccountable participants indeed used the category as a basis for drawing inferences.

By contrast, accountable participants did not appear to be influenced by the social category or, as an alternative interpretation, accountable participants suppressed the use of categories due to their lack of social acceptability whereas their ease of use was too appealing to resist for unaccountable participants. There was no shift toward less favorable performance assessments due to immigration status. Furthermore, the analysis revealed a strong decrease in the total number of inferences and the number of category-consistent inferences as compared to unaccountable participants.

In sum, then, our results lend credence to the hypothesis that teachers' judgments of student performance and their attributions of students' characteristics are biased by activated social categories or more correctly reflect the information given about the student depending on accountability. Thus, our findings supported Hypotheses 1 and 2. As far as we know, this is the first demonstration that knowledge about the determinants of category-based versus attribute-based judgment formation can be applied to the domain of educational achievement judgments. A number of questions about this effect, however, remain. First, accountable and unaccountable participants may have differed in their use of a category for the judgment task, but not in their activation of a category. It is thus not clear whether participants in both accountability conditions may have activated a category, but the accountable participants may have been more cautious about using the category for the judgment. However, from our theoretical background, it can be assumed that unaccountable teachers activate categories during their encoding of the information, whereas accountable teachers direct their attention to individual attributes of students, which may impede category activation. Thus, accountable and unaccountable participants should differ in category activation while encoding the information. Second, it is not clear whether accountability still exerts an effect when categories of professional pedagogical content knowledge are used. In Study 1, social categories from everyday knowledge about persons were used. It has been well-documented that this type of person category affects teachers' attention, memory and attributions for students' achievements under certain conditions (Krolak-Schwerdt et al. 2009; Reyna 2000; Weiner 2000). However, experienced teachers have developed a pedagogical content knowledge base (cf. Shulman 1986), which comprises typologies of specific student categories (Hofer 1981). Examples of these student categories are "No. 1 in class" or "hyperactive child." One may well question

whether teachers' processing of information from everyday knowledge is different from their processing on a professional knowledge base. Study 2 addressed these questions.

3 Study 2

To investigate the question of whether accountable and unaccountable participants differ in category activation while encoding student information (Hypothesis 3), we adopted a paradigm developed by Albrecht and O'Brien (1993). In this paradigm, reading times for person information are used as dependent variables. After making a category available, the reader of a person description receives information about a fictitious person in which one sentence (the critical experimental sentence) of the description contradicts the category (or inferences that might be drawn from the category) or else the critical sentence is consistent with the contents of the category. An increase in reading time for the contradictory statement as compared to the consistent statement would show that the category was activated while the participant read the person description because it is only possible to notice the contradiction when the category is cognitively activated at the moment in which the reader encounters the description. Noticing the contradiction should induce comprehension difficulties, which in turn should increase reading time (Albrecht and O'Brien 1993; Gernsbacher et al. 1992).

3.1 Methods

3.1.1 Participants and design

Twenty-four German secondary school teachers (12 male, 12 female) participated in the study. On average, the participants were 52.8 years old (SD = 6.4) and had 23.8 years (SD = 8.5) of experience in their profession. They were randomly assigned to the conditions of a 2 (accountability: accountable vs. unaccountable) × 2 (consistency: contradictory sentence vs. consistent sentence) mixed factorial design.

3.1.2 Materials and procedure

Participants, on their arrival in the laboratory, were given one of the accountability instructions from Experiment 1. Then they were handed the case materials, which consisted of the behavior and achievement descriptions of two students. Instructions and case materials were presented on a computer screen. In each experimental condition and for each student description, a student category was activated. One sentence of the description either contradicted the student category or was consistent with the activated category. In one condition participants received first the description with the consistent sentence followed by the description with the contradictory statement. In a second condition the order of descriptions was reversed. Each description was presented sentence by sentence according to the self-paced reading time method (Haberlandt 1994). Participants had to read each sentence, which was presented in a window

in the center of a computer screen, at their own pace. By pressing a key after reading each sentence, the sentence vanished and an asterisk appeared in the center of the screen. The participant had to fixate on the asterisk to prevent large eye movements between presentations of successive sentences and to ensure that the participant's gaze was at the center when the next sentence was presented. The asterisk vanished after 50 ms, and the next sentence appeared in the window at the center of the screen. The time between the onset of each sentence presentation and the following key press was defined as the reading time for the sentence (see Haberlandt 1994). After completing the reading task, participants were debriefed, thanked for their participation, and dismissed.

To create the descriptions, typologies of student categories were used; these typologies had been demonstrated to be part of the pedagogical content knowledge base of experienced teachers. From these, two student categories (which were "No. 1 in class" and "introverted-withdrawn student") were selected with student characteristics that teachers highly agreed upon and that have been consensually used by teachers (Hofer 1981; Hörstermann and Krolak-Schwerdt 2012). Each description consisted of 20 statements each pertaining to a characteristic of the selected student category and five statements that were neutral with respect to the student category. The statements of each description were presented in the following order. First, 15 statements that pertained to the student category were presented in order to activate the selected category. These were followed by four neutral statements. The 20th sentence of the description was the critical experimental statement, which either contradicted the category or was consistent with the category. Afterwards, four additional statements pertaining to the category and one additional neutral statement were presented.

In one experimental condition the contradictory experimental sentence appeared in the description of "No. 1 in class" and the consistent experimental sentence appeared in the description of the "introverted-withdrawn student" (termed set 1 in the following), whereas in another condition the allocation of the experimental sentences to the descriptions was reversed (termed set 2 in the following). Of course, the contradictory experimental sentence in set 1 opposed the student category "No. 1 in class", whereas the contradictory experimental sentence in set 2 opposed the student category "introverted-withdrawn student". Participants of the study were randomly assigned to the conditions of order of presentation and set 1 and set 2.

Only the reading times for the experimental sentence were of interest for the study. Of course, participants were not made aware of the type of information that was of interest.

3.1.3 Preprocessing of the reading-time data

In order to eliminate effects of reading times for the experimental sentence that were due to differences in sentence length, reading times were normalized in the following way: For each participant and each description, the time for the experimental sentence was divided by the number of syllables in the sentence (see, e.g., Haberlandt 1994). Thus, the dependent variable can be interpreted as reading time per syllable and was measured in milliseconds.

Table 3 Mean reading times per syllable (in milliseconds) for the experimental sentence as a function of accountability and consistency of statement (Study 2)

Accountability	Consistency of statement	
	Consistent	Contradictory
Accountable	174.82	154.69
Unaccountable	131.75	260.24

3.2 Results and discussion

To address the predictions regarding the activation of the student category while teachers passed through the student description, a 2 (accountability: accountable vs. unaccountable) × 2 (order of presentation: description with consistent sentence first vs. description with contradictory sentence first) × 2 (set: set 1 vs. set 2) × 2 (consistency: contradictory sentence vs. consistent sentence) ANOVA with repeated measures on the last factor was conducted on the reading times for the experimental sentence. The interaction of accountability and consistency was significant, $F(1, 16) = 4.81$, $\eta^2 = .23$, $p = .04$. No other effects were significant, all $Fs < 2.5$. As can be seen in Table 3, the pattern of means confirmed our predictions. Planned contrasts showed that in the accountable condition, reading times for the contradictory and consistent statements did not differ, $t(11) = -.46$, $p = .65$, *ns*. However, there was a strong effect of the consistency of the statement in the unaccountable condition, $t(11) = 2.91$, $d = .84$, $p = .01$. It took participants twice as long to read the contradictory statement than to read the consistent sentence.

From these findings, it may be concluded that unaccountable participants had difficulty comprehending the contradictory statement. In the framework of our experiment, comprehension difficulties would occur only when participants tried to map the critical experimental sentence onto the previously activated category. This in turn implies that unaccountable participants had activated a student category during their encoding of the information presented in the description. By contrast, from the findings in the accountable condition, it may be concluded that participants did not realize the contradiction between the experimental sentence and the category-related information of the description. This result suggests that accountable participants had not activated a student category during the encoding of the description. In sum, our findings supported Hypothesis 3.

Having demonstrated that unaccountable teachers activated a student category during passing through a student description, whereas there were no indices of category activation for accountable teachers, further questions on the processing of student characteristics and the representation of these characteristics in teachers' memory remain. From Experiment 2, there was no evidence regarding whether accountable and unaccountable teachers encode and represent the whole description in memory differently. It is thus unclear whether an activated category exerts an influence on the process of organization, and storage of the student information that was given. From our theoretical background, it may be expected that unaccountable teachers pay little

attention to the individual details of student information if they have a category at hand, whereas accountable teachers direct their attention to the details of the description irrespective of category availability. As attention determines which information is encoded and stored in memory, it may be predicted that unaccountable teachers' memory of students' attributes will be biased by an activated category, whereas accountable teachers' memory will not be influenced by a category. Study 3 addressed this question.

4 Study 3

To investigate the question of whether teachers' memory of students' attributes is differentially influenced by an activated category depending on accountability (Hypothesis 4), we used a free recall task. Participants received two student descriptions, one pertaining to a student category and the other comprising individual, that is, category-unrelated student characteristics. To assess memory, participants had to reproduce as much of the presented student information as they could remember.

4.1 Methods

4.1.1 Participants and design

Forty German secondary school teachers (17 male, 23 female) participated in the study. On average, the participants were 45.2 years old (SD = 9.6) and had 18.1 years (SD = 9.8) of experience in their profession. They were randomly assigned to the conditions of a 2 (accountability: accountable vs. unaccountable) $\times$ 2 (type of description: category-related vs. individual) mixed factorial design with repeated measures on the type of description.

4.1.2 Materials and procedure

Participants, upon their arrival in the laboratory, were given one of the accountability instructions from Study 1. Then they received two student descriptions, one pertaining to a student category (termed the category-related description in the following) and the other one describing individual characteristics that were category-unrelated (termed the individual description in the following). Each description was presented on the computer screen. Half of the participants received the category-related description first and the other half received the individual description first. Participants were randomly assigned to the order of presentation of the two descriptions. The presentation phase of the experiment was followed by a short interpolated task to interfere with participants' short-term memory in which participants had to fill out a questionnaire on their biographical data. Afterwards, participants were asked to recall as much of the original information as they could.

It should be noted that Studies 1 and 3 used research paradigms that were established in different domains of research and that yield other types of dependent variables, and

thus, Study 3 may not be considered to be a replication of Study 1. In the domain of text comprehension, summarization in a person's own words is used to analyze inferences from person descriptions as has been noted above, whereas in social cognition research, the free recall task is used to analyze correct reproductions of the given information and memory intrusions. Correct reproductions are pieces of information repeated verbatim and thus do not go beyond the given information as inferences do. Memory intrusions may comprise inferences, but they do not necessarily correspond to inferences in every case as they also include any other type of cues that falsely come into mind while working on the recall task.

To obtain the category-related student description, we proceeded as we had in Study 2. Thus, the category-related description consisted of 20 statements each pertaining to a characteristic of a selected student category (which was "hyperactive child" in this experiment) and four statements that were neutral with respect to the student category. The individual student description comprised 24 statements, which were neutral with respect to category activation.

A prestudy had the aim of investigating whether the descriptions represented the intended student cases. Ten secondary school teachers with at least 10 years of professional school experience had to estimate the ease of categorizing each student description on a 3-point rating scale ranging from 1 (*easy*) to 3 (*difficult*), and they had to specify a student category that fit the description. For the individual description, all participants answered that they were not able to specify any category. On average, the description was rated 2.89 and was thus judged difficult to categorize. For the category-related description, eight participants mentioned the intended category "hyperactive child." On average, the description was rated 1.12, and was therefore judged easy to categorize.

4.1.3 Dependent measures and coding

Participants' recall protocols were scored by two independent judges, blind to the experimental condition, for correct reproductions and intrusions. Reproductions were scored as having been correctly remembered if a statement from the case report about the student was repeated verbatim or if a synonymous formulation was created by the participant, and as intrusions in any other case. In the second step, intrusions were coded with regard to whether they pertained to a student category. An intrusion was judged as pertaining to the corresponding category if it was a verbatim or synonymous formulation of an attribute of the category although it was not part of the description in the presentation phase of the experiment. The typologies of teachers' pedagogical content knowledge outlined above were used as a basis for these judgments. Interrater reliability (Cohen's kappa) was $\kappa = .89$.

Relative correct recall rates were calculated by dividing the number of correctly recalled statements by the total number of statements presented in the description, that is, 24 statements, and relative numbers of intrusions were computed by dividing the number of intrusions by the total number of productions, that is, the sum of correctly recalled statements and the number of intrusions for each description and each participant (see Murphy and Puff 1982, for this procedure).

Table 4 Mean correct recall rates as a function of accountability and category availability (Study 3)

Accountability	Type of description	
	Category-related	Individual
Accountable	0.32	0.33
Unaccountable	0.21	0.35

Relative correct recall rates were calculated by dividing the number of correctly recalled statements by the total number of statements presented in the description, that is, 24 statements

Table 5 Mean relative number of intrusions as a function of accountability and category availability (Study 3)

Accountability	Type of description	
	Category-related	Individual
Accountable	0.13	0.15
Unaccountable	0.51	0.19

The relative number of intrusions was calculated by dividing the total number of intrusions by the total number of reproductions (i.e., intrusions and correctly recalled statements)

4.2 Results and discussion

Two dependent variables were of primary concern: The number of correctly recalled statements, and the number of intrusions. From the various ways to treat reproductions and intrusions, their separate analysis was selected in order to avoid any confounding (see Murphy and Puff 1982). Results for each of the two variables will be discussed below.

With respect to teachers' memory of students' characteristics, we expected that the number of correctly recalled statements would be less for the category-related compared to the individual description for unaccountable participants, whereas there would be no difference for accountable participants. Correspondingly, a 2 (accountability) × 2 (order of presentation) × 2 (type of description) ANOVA was conducted on the relative number of correctly recalled attributes. There was a main effect of type of description, $F(1, 36) = 7.06$, $\eta^2 = .17$, $p = .01$, with a higher recall rate for the individual description ($M = 0.34$) than for the category-related case report ($M = 0.26$). Furthermore, the interaction of accountability and type of description was significant, $F(1, 36) = 4.20$, $\eta^2 = .11$, $p = .05$. No other effects were significant. Table 4 shows the mean recall rates, which were consistent with predictions. Planned contrasts confirmed that unaccountable participants reproduced fewer statements correctly for the category-related case compared to the individual description, $t(19) = -4.91, d = 1.10, p = .0009$, whereas accountable participants showed no differences in recall rates between these two descriptions, $t(19) = -.39, p = .71$.

To further analyze our prediction that unaccountable teachers' memory would be biased by an activated category, whereas accountable teachers' memory would more accurately reflect the information given about the student, a 2 (accountability) × 2

Table 6 Mean relative number of category-consistent intrusions as a function of accountability and category availability (Study 3)

Accountability	Type of description	
	Category-related	Individual
Accountable	0.05	0.06
Unaccountable	0.32	0.10

The relative number of category-consistent intrusions was calculated by dividing the total number of category-consistent intrusions by the total number of reproductions (i.e., intrusions and correctly recalled statements)

(order of presentation) × 2 (type of description) ANOVA was performed on the relative number of intrusions. There was a main effect of accountability, $F(1, 36) = 14.19$, $\eta^2 = .31$, $p = .001$, with a higher number of intrusions for unaccountable participants ($M = 0.35$) than for accountable participants ($M = 0.14$) and a main effect of type of description, $F(1, 36) = 18.05$, $\eta^2 = .36$, $p = .001$, with a higher number of intrusions for the category-related description ($M = 0.32$) than for the individual description ($M = 0.17$). Most interestingly for our research question, there was a strong interaction effect of accountability and type of description, $F(1, 36) = 22.06$, $\eta^2 = .41$, $p = .001$. No other effects were significant, all $Fs < 1$. As the mean number of intrusions displayed in Table 5 show, the results were consistent with predictions. There appeared a strong increase in the number of intrusions for the category-related description compared to the individual case report for unaccountable participants, $t(19) = 5.58$, $d = 1.25$, $p = .0002$, whereas accountable participants did not exhibit a difference in intrusions between the two student descriptions, $t(19) = -.801$, $p = .44$.

In order to investigate the question of whether intrusions indeed pertained to the activated category, we analyzed the relative number of category-consistent intrusions as a final step. A 2 (accountability) × 2 (order of presentation) × 2 (type of description) ANOVA revealed the following results. There was a main effect of accountability, $F(1, 36) = 15.75$, $\eta^2 = .33$, $p = .001$, with a higher number of category-consistent intrusions for unaccountable participants ($M = 0.21$) than for accountable participants ($M = 0.06$) and a main effect of type of description, $F(1, 36) = 15.11$, $\eta^2 = .32$, $p = .001$, with a higher number of category-consistent intrusions for the category-related description ($M = 0.18$) than for the individual description ($M = 0.08$). Furthermore, there was a strong interaction effect of accountability and type of description, $F(1, 36) = 18.85$, $\eta^2 = .37$, $p = .001$. Table 6 shows the mean number of category-consistent intrusions. For unaccountable participants, there was a strong increase in category-consistent intrusions for the category-related case report compared to the individual description, $t(19) = 5.14$, $d = 1.15$, $p = .0005$, whereas no such increase appeared for accountable participants, $t(19) = -.74$, $p = .47$. A comparison of the number of category-related intrusions to the total number of intrusions for the category-related description revealed that 69 % of the intrusions for unaccountable participants pertained to the activated category.

In sum, the findings from Study 3 strongly supported Hypothesis 4. As the results from the free recall data demonstrated, unaccountable teachers' memory was impaired

by the presence of a student category, whereas there was no such effect in every other condition. The analysis of intrusions and, more specifically, category-consistent intrusions, finally demonstrated that for unaccountable participants, the activated category caused the memory bias, whereas there was no bias for accountable teachers, but a relatively more correct memory for the given information.

5 General discussion

The present studies examined whether the accountability of teachers has an influence on their processing of students' achievements and their assessments of student performance. Building on research on social judgment formation (Fiske and Neuberg 1990; Lerner and Tetlock 2003) we predicted that unaccountable teachers would activate social categories and use them for the assessment, whereas accountable teachers' attention would be directed to individual attributes of students. Using secondary school teachers as participants, three studies supported our hypotheses. First, we found in Study 1 that unaccountable teachers' assessments of student performance were biased by activated social categories and that unaccountable teachers generated inferences about students on the basis of the activated categories. By contrast, accountable teachers more accurately took into account the information given about a student in their assessments and a comparatively low number of inferences was observed. Second, we found in Study 2 that unaccountable teachers activated social categories about students during the encoding of the student information, whereas accountable teachers did not activate a category during the encoding phase of information processing. Finally, in Study 3 we found that unaccountable teachers' memory was influenced by social categories, but there were no such biases on the base of categories in the memory of accountable teachers.

From our findings it may be concluded that accountability differentially shifts teachers' processing of student information and their assessments of student performance to more category-based or attribute-based strategies. Thus, the dual process theories underlying the present research provide a solid base for explaining teachers' information processing and judgment formation. Low accountability induces category-based processing with attention, memory and judgment being affected by social categories, whereas high accountability directs attention to the individual information given about a student with memory and judgment being unaffected by categories. Thus, teachers are "motivated tacticians" in their professional domain who select from at least two processing strategies in response to the actual demand. As far as we know, this is the first demonstration that knowledge about the determinants of category-based versus attribute-based judgment formation can be applied to the domain of educational achievement judgments. In addition, the present research sheds light on the cognitive processes that underlie the variations in the quality of teachers' assessments by demonstrating that accountability influences early phases in the processing of student information, that is, attention and memory. This in turn may also constitute the cognitive mechanisms of relatively more biased or accurate judgment formation in the educational domain.

Our findings add to a large body of evidence demonstrating that teachers' expectations about students' achievements and their corresponding assessments of student performance may be influenced by activated student categories (Andrews et al. 1997; Parks and Kennedy 2007; Pigott and Cowen 2000). However, it also extends knowledge on teachers' judgment formation by showing conditions in which teachers' assessments are relatively free of biases (that is, high accountability). Thus, conceptualizing this type of assessment as a social judgment and using dual process theories allowed us to formulate new hypotheses on conditions which moderate the quality of educational achievement assessments.

Our results on the impact of accountability on teachers' assessments of student performance fit into Tetlock's research (Tetlock 2005; Tetlock and Lerner 1999) on professional judgment formation. According to Tetlock (2005) judgments are more accurate if professional forecasters (e.g. political experts, engineers, or teachers) are held systematically accountable. By contrast, a "regime of close-to-zero accountability" (Tetlock 2005, p. 235) can lead to an accuracy that is worse than flipping a coin. Even if accountability by itself is no panacea, it has an impact on cognitive processing and can lead to a more complex, self-critical and effort-demanding style of thinking (Tetlock and Lerner 1999). So teachers like other forecasters should be held responsible for their judgments, e.g. by ensuring that they expect to explain, justify, or excuse their judgments. From this line of research as well as research on decision making of physicians and clinical psychologists (Dawes 1998; Swets et al. 2000) attribute-based processing is the competent mode for highly consequential decisions because it is characterized by the systematic collection of pieces of information and their thoughtful integration into a judgment. Accountability seems to trigger this type of judgment.

On the other hand, less consequential assessments might still qualify as "good enough" even if they are based on category-related processing (Fiske 1993). Examples are decisions during class such as calling on a particular student. In these situations, attribute-based processing might be overly demanding, whereas category-based processing might reduce cognitive load during judgment formation and thus help teaching in the class. In other words, although attribute-based processing appears as the competent mode for consequential assessments, category-based judgment formation might serve the capability of acting with a reasonable quality for less consequential decisions.

One major objection against our studies might concern the instructions we used to vary accountability. The instructions varied accuracy and significance of the assessment at the same time. In the *accountable* condition participants were informed that they should recommend a secondary school track for students which constitutes a highly consequential judgment, whereas participants in the *unaccountable* condition received no such information. Thus, either accuracy or significance or both might have been the relevant determinants for our findings. It should be noted that it was not the aim of our studies to separately analyze the impact of these two determinants on judgment formation. Rather, our intention was to identify those tasks in the everyday professional life of teachers which differ in accountability and to operationalize the tasks in the experiments through instruction. As has been outlined at the outset, we conducted two prestudies (Prestudy 1 and 2) to guarantee the external validity of the instructions. Results of the prestudies showed that our instructions correspond to professional tasks

of teachers in the school setting and indeed differed in accountability. Furthermore, the two instructions used in our studies may be differentiated from each other as inducing proximal versus distal goals (see, e.g., Gollwitzer and Moskowitz 1996). These two types of goals have been shown to differentially affect social cognition and they inherently comprise accuracy, significance and level of abstraction. Thus, accuracy and significance may not be separated from each other beyond an experimental rationale.

Another objection against our studies might concern that we also varied the student descriptions. Across the studies of the present investigation we selected different student categories as part of the materials and investigated the validity of the student descriptions based on these materials in a number of prestudies. Materials were varied in order to test whether our results generalize across different student descriptions or represent effects due to materials. The consistency of results implies that our findings generalize across different materials.

The type of category-based processing that we investigated in our studies may be conceptualized as one example of heuristic decision making. Other examples of using heuristics that might apply to the educational domain are anchoring and adjustment. Anchoring effects occur when a preceding assessment has an influence on subsequent assessments (Mussweiler and Strack 1999; Tversky and Kahnemann 1974). In the educational domain anchoring effects might occur in assessing several class tests where the first test may yield the anchor for the subsequent tests. In our current studies we investigate whether teachers use the anchoring heuristic in achievement assessments. Another research aim concerns the question whether anchoring effects are moderated by accountability in the same way as category-based processing. In this case, the theoretical foundation of category-based processes might be broadened to the more general conceptualization of heuristic thinking and accountability might be considered as a general moderator of heuristic thinking. Clearly, the supposed generality of these mechanisms remains to be tested empirically.

References

Albrecht, J. E., & O'Brien, E. J. (1993). Updating a mental model: Maintaining both local and global coherence. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *19*(5), 1061–1070. doi:10.1037/0278-7393.19.5.1061.

Alpert, B., & Bechar, S. (2008). School organisational efforts in search for alternatives to ability grouping. *Teaching and Teacher Education*, *24*(6), 1599–1612. doi:10.1016/j.tate.2008.02.023.

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

Babad, E. (1993). Teachers' differential behavior. *Educational Psychology Review*, *5*(4), 347–376. doi:10.1007/BF01320223.

Baumert, J., Artelt, C., Klieme, E., Neubrand, M., Prenzel, M., Schiefele, U., et al. (Eds.). (2002). *PISA 2000—Die Länder der Bundesrepublik Deutschland im Vergleich (PISA 2000—The states of the Federal Republik of Germany in comparison)*. Opladen, Germany: Leske & Budrich.

Baumert, J., Klieme, E., Neubrand, M., Prenzel, M., Schiefele, U., Schneider, W., et al. (Eds.). (2001). *PISA 2000: Basiskompetenzen von Schülerinnen und Schülern im internationalen Vergleich (PISA 2000: Basic skills of students in an international comparison)*. Opladen, Germany: Leske & Budrich.

BMBF. (2001). *TIMSS—Impulse für Schule und Unterricht: Forschungsbefunde, Reforminitiativen, Praxisberichte und Video-Befunde (TIMSS—Impulses for school and teaching: Research findings, reform initia-*

tives, practical experience reports, and video evidences). Bonn, Germany: Author. Retrieved December 22, 2011, from www.bmbf.de/pub/timss.pdf.

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

Brewer, M. B. (1988). A dual process model of impression formation. In T. K. Srull & R. S. Wyer (Eds.), *Advances in social cognition* (Vol. 1, pp. 1–36). Hillsdale, NJ: Lawrence Erlbaum Associates.

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

Chaiken, S., & Ledgerwood, A. (2011). A theory of heuristic and systematic information processing. In P. A. M. Van Lange, A. W. Kruglanski, & E. T. Higgins (Eds.), *Handbook of theories of social cognition* (Vol. 1, pp. 246–266). London: Sage.

Chen, S., & Chaiken, S. (1999). The heuristic-systematic model in its broader context. In S. Chaiken & Y. Trope (Eds.), *Dual-process theories in social psychology* (pp. 73–96). New York, NY: Guilford Press.

Dawes, R. M. (1998). Behavioral decision making and judgment. In D. T. Gilbert, S. T. Fiske, & G. Lindzey (Eds.), *The handbook of social psychology* (4th ed., Vol. 1, pp. 497–548). New York, NY: McGraw-Hill.

Fiske, S. T. (1993). Social cognition and social perception. *Annual Review of Psychology*, *44*, 155–194. doi:10.1146/annurev.ps.44.020193.001103.

Fiske, S. T. (2011). The continuum model and the stereotype content model. In P. A. M. Van Lange, A. W. Kruglanski, & E. T. Higgins (Eds.), *Handbook of theories of social cognition* (Vol. 1, pp. 267–288). London: Sage.

Fiske, S. T., Lin, M., & Neuberg, S. L. (1999). The continuum model: Ten years later. In S. Chaiken & Y. Trope (Eds.), *Dual-process theories in social psychology* (pp. 231–254). New York, NY: Guilford Press.

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

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

Gernsbacher, M. A. (1996). Coherence cues mapping during comprehension. In J. Costermans & M. Fayol (Eds.), *Processing interclausal relationships in the production and comprehension of text* (pp. 3–21). Hillsdale, NJ: Lawrence Erlbaum Associates.

Gernsbacher, M. A., Goldsmith, H. H., & Robertson, R. R. W. (1992). Do readers mentally represent characters' emotional states? *Cognition and Emotion*, *6*(2), 89–111. doi:10.1080/02699939208411061.

Gilbert, D. T. (1998). Ordinary personology. In D. T. Gilbert, S. T. Fiske, & G. Lindzey (Eds.), *The handbook of social psychology* (4th ed., Vol. 2, pp. 89–150). Boston, MA: McGraw-Hill.

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

Goldenberg, C. (1992). The limits of expectations: A case for case knowledge about teacher expectancy effects. *American Educational Research Journal*, *29*(3), 517–544. doi:10.3102/00028312029003517.

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

Good, T. L., & Nichols, S. L. (2001). Expectancy effects in the classroom: A special focus on improving the reading performance of minority students in first-grade classrooms. *Educational Psychologist*, *36*(2), 113–126. doi:10.1207/S15326985EP3602_6.

Graham, S. (1990). On communicating low ability in the classroom. In S. Graham & V. Folkes (Eds.), *Attribution theory: Applications to achievement, mental health, and interpersonal conflict* (pp. 17–36). Hillsdale, NJ: Lawrence Erlbaum Associates.

Haberlandt, K. (1994). Methods in reading research. In M. A. Gernsbacher (Ed.), *Handbook of psycholinguistics* (pp. 1–31). San Diego, CA: Academic Press.

Hofer, M. (1981). Schülergruppierungen in Urteil und Verhalten des Lehrers (Student grouping in judgment and behavior of the teacher). In M. Hofer (Ed.), *Informationsverarbeitung und Entscheidungsverhalten von Lehrern: Beiträge zu einer Handlungstheorie des Unterrichtens* (pp. 192–221). München: Urban & Schwarzenberg.

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

Hörstermann, T., & Krolak-Schwerdt, S. (2012). Teachers' typology of student categories. A cluster analytic study. In W. Gaul, A. Geyer-Schulz, & L. Schmidt-Thieme (Eds.), *Studies in classification, data analysis and knowledge organization, Vol. 43: Challenges at the interface of data analysis, computer science, and optimization* (pp. 547–555). Berlin: Springer.

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

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. doi:10.1207/s15327957pspr0902_3.

Kelley, H. H. (1967). Attribution theory in social psychology. In D. Levine (Ed.), *Nebraska symposium on motivation* (Vol. 15, pp. 192–238). Lincoln, NE: University of Nebraska Press.

Krolak-Schwerdt, S., Böhmer, M., & Gräsel, C. (2009). Verarbeitung von schülerbezogener Information als zielgeleiteter Prozess: Der Lehrer als, flexibler Denker (Goal-directed processing of students' attributes: The teacher as "flexible thinker"). *Zeitschrift für Pädagogische Psychologie*, *23*(3–4), 175–186. doi:10.1024/1010-0652.23.34.175.

Langer, A., Langer, H., & Theimer, H. (1993). *Lehrer beobachten und beurteilen Schüler (Teachers observe and assess students)*. München: Oldenbourg.

Lerner, J. S., & Tetlock, P. E. (1994). Accountability and social cognition. In V. S. Ramachandran (Ed.), *Encyclopedia of human behavior* (Vol. 1, pp. 3098–3121). San Diego, CA: Academic Press.

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

Marsh, H. W. (1987). The big-fish-little-pond effect on academic self-concept. *Journal of Educational Psychology*, *79*(3), 280–295. doi:10.1037/0022-0663.79.3.280.

Marsh, H. W., Kong, C.-H., & Hau, K.-T. (2000). Longitudinal multilevel models of the big-fish-little-pond effect on academic self-concept: Counterbalancing contrast and reflected-glory effects in Hong Kong schools. *Journal of Personality and Social Psychology*, *78*(2), 337–349.

Murphy, M. D., & Puff, C. R. (1982). Free recall: Basic methodology and analyses. In C. R. Puff (Ed.), *Handbook of research methods in human memory and cognition* (pp. 99–128). New York, NY: Academic Press.

Mussweiler, T., & Strack, F. (1999). Hypothesis-consistent testing and semantic priming in the anchoring paradigm: A selective accessibility model. *Journal of Experimental Social Psychology*, *35*(2), 136–164. doi:10.1006/jesp.1998.1364.

Ormrod, J. E. (2006). *Educational psychology: Developing learners* (5th ed.). Upper Saddle River: Pearson Prentice Hall.

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. doi:10.1177/0021934705285955.

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

Reyna, C. (2000). Lazy, dumb or industrious. When stereotypes convey attribution information in the classroom. *Educational Psychology Review*, *12*(1), 85–110. doi:10.1023/A:1009037101170.

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

Shepard, L. A. (2006). Classroom assessment. In R. L. Brennan (Ed.), *Educational measurement* (4th ed., pp. 623–646). Westport, CT: American Council on Education/Praeger.

Shulman, L. S. (1986). Those who understand: Knowledge growth in teaching. *Educational Researcher*, *15*(2), 4–14. doi:10.3102/0013189X015002004.

Südkamp, A., Kaiser, J., Möller, J. (2012, March 26). Accuracy of teachers' judgments of students' academic achievement: A meta-analysis. *Journal of Educational Psychology*, Advance online publication. doi:10.1037/a0027627.

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

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

Tetlock, P. E. (2005). *Expert political judgment: How good is it? How can we know?*. Princeton, NJ: Princeton University Press.

Tetlock, P. E., & Lerner, J. S. (1999). The social contingency model: Identifying empirical and normative boundary conditions on the error-and-bias portrait of human nature. In S. Chaiken & Y. Trope (Eds.), *Dual-process theories in social psychology* (pp. 571–585). New York, NY: Guilford Press.

Tversky, A., & Kahnemann, D. (1974). Judgment under uncertainty: Heuristics and biases. *Science*, *185*(4157), 1124–1131. doi:10.1126/science.185.4157.1124.

Van Ophuysen, S. (2006). Vergleich diagnostischer Entscheidungen von Novizen und Experten am Beispiel der Schullaufbahnempfehlung (Comparison of diagnostic decisions between novices and experts: The example of school career recommendations). *Zeitschrift für Entwicklungspsychologie und Pädagogische Psychologie*, *38*(4), 154–161.

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

Wintermantel, M., & Christmann, U. (1983). Person description: Some empirical findings concerning the production and reproduction of a specific text-type. In G. Rickheit & M. Bock (Eds.), *Psycholinguistic studies in language processing* (pp. 137–151). Berlin: De Gruyter.

Wintermantel, M., & Krolak-Schwerdt, S. (2002). Eindrucksbildung aus Personbeschreibungen (Impression formation from person descriptions). *Zeitschrift für Sozialpsychologie*, *33*(1), 45–64. doi:10.1024//0044-3514.33.1.45.

Author Biographies

Sabine Krolak-Schwerdt, Ph.D. is Professor of Educational Measurement at the University of Luxembourg. She heads the research area Professionalization of Actors in Education Fields at the Faculty of Language and Literature, Humanities, Arts and Education. Her research interests focus on Educational Psychology, Social Cognition, and Multivariate Methods.

Matthias Böhmer, Ph.D. is Research Assistant at the University of Luxembourg. He is a member of the Faculty of Language and Literature, Humanities, Arts and Education. His research focuses on Educational and Clinical Expertise, REBT, and Social Cognition.

Cornelia Gräsel, Ph.D. is Professor of Learning and Instruction at the University of Wuppertal. She heads the Institute of Educational Research at the School of Education. Her research interests focus on Teacher Education, Teacher Competencies, and School Transitions.