Provided for non-commercial research and education use.
Not for reproduction, distribution or commercial use.

Volume 49 Issues 2–3 2010 ISSN 0883-0355

ELSEVIER

International Journal of

Educational Research

Individual articles

Editor: Neil Mercer

Deputy Editor: Andreas Stylianides

Available online at

ScienceDirect

www.sciencedirect.com

This article appeared in a journal published by Elsevier. The attached copy is furnished to the author for internal non-commercial research and education use, including for instruction at the authors institution and sharing with colleagues.

Other uses, including reproduction and distribution, or selling or licensing copies, or posting to personal, institutional or third party websites are prohibited.

In most cases authors are permitted to post their version of the article (e.g. in Word or Tex form) to their personal website or institutional repository. Authors requiring further information regarding Elsevier's archiving and manuscript policies are encouraged to visit:

http://www.elsevier.com/copyright

Author's personal copy

International Journal of Educational Research 49 (2010) 78–91

Contents lists available at ScienceDirect

International Journal of Educational Research

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

Does immigration background matter? How teachers' predictions of students' performance relate to student background[☆]

Axinja Hachfeld [a,*], Yvonne Anders [b], Sascha Schroeder [a,d], Petra Stanat [c,e], Mareike Kunter [a,f]

[a] *Max Planck Institute for Human Development, Lentzeallee 94, 14195 Berlin, Germany*
[b] *Otto Friedrich University Bamberg, Postfach 1549, 96045 Bamberg, Germany*
[c] *Free University Berlin, Habelschwerdter Allee 45, 14195 Berlin, Germany*
[d] *University of Kassel, Arnold-Bode-Str. 10, 34109 Kassel, Germany*
[e] *Institute for Educational Progress, Unter den Linden 6, 10099 Berlin, Germany*
[f] *Johann Wolfgang Goethe University, Postfach 125, 60054 Frankfurt am Main, Germany*

ARTICLE INFO

Article history:
Received 7 January 2010
Received in revised form 23 August 2010
Accepted 21 September 2010

Keywords:
Immigrant students
Teacher predictions
Prediction accuracy
Bilingual students
International education

ABSTRACT

Accurate teacher evaluations of student performance are crucial for effective teaching. This study examined whether students' immigration and language background affect teachers' evaluations. Multilevel analyses tested whether teachers overestimate the performance of immigrant relative to that of non-immigrant students. As part of the German PISA 2003 assessment, 305 teachers predicted the performance of seven of their students on two mathematics problems of different linguistic complexity. Results revealed an interaction effect of students' language background and linguistic complexity of the problem on teachers' predictions. Teachers overestimated the performance of bilingual students more than the performance of monolingual immigrant or non-immigrant students on a linguistically complex problem. Teachers need to consider both students' language background and linguistic demands of the material used to appropriately support bilingual students.

© 2010 Elsevier Ltd. All rights reserved.

1. Introduction

Low academic performance of minority or immigrant students has been a stable phenomenon for decades in many Western countries, as documented by large-scale international assessments such as the Programme for International Student Assessment (PISA). Results from PISA indicate that immigrant students often perform at levels significantly lower than non-immigrant students, even though they are often motivated learners (Organisation for Economic Cooperation and Development [OECD], 2004; Stanat & Christensen, 2006). On a societal level, the performance gap between minority and majority students poses many problems, and researchers from different fields and countries have discussed the role education should play in days of rising cultural pluralism (cp. Phillips & Callan, 2001). The performance gap has been found to be especially pronounced in Germany, where immigrant students are overrepresented in groups at risk for academic failure and underrepresented in college-bound school tracks (Organisation for Economic Cooperation and Development [OECD],

[☆] The research reported in this article is based on data from the Professional Competence of Teachers, Cognitively Activating Instruction, and the Development of Students' Mathematical Literacy study (COACTIV) directed by Jürgen Baumert (Max Planck Institute for Human Development, Berlin, Germany), Werner Blum (Kassel University, Germany), and Michael Neubrand (Carl von Ossietzky University, Oldenburg, Germany). The project is funded by the German Research Foundation (DFG; BA1461/2-2) as part of its priority Program on School Quality (BIQUA).

[*] Corresponding author. Tel.: +49 30 82406 671; fax: +49 30 82406 490.
E-mail addresses: hachfeld@mpib-berlin.mpg.de (A. Hachfeld), Yvonne.anders@uni-bamberg.de (Y. Anders), Sascha.schroeder@mpib-berlin.mpg.de (S. Schroeder), petra.stanat@iqb.hu-berlin.de (P. Stanat), kunter@paed.psych.uni-frankfurt.de (M. Kunter).

0883-0355/$ – see front matter © 2010 Elsevier Ltd. All rights reserved.
doi:10.1016/j.ijer.2010.09.002

Author's personal copy

2004; Stanat & Christensen, 2006). At the same time, the number of students who are foreign-born or have foreign-born parents is growing rapidly. In the following, these students will be referred to as immigrant students.

Since the publication of the first PISA results, teachers have been blamed for failing to provide immigrant students with proper support, and thereby perpetuating the achievement gap. Teachers have been held responsible for the achievement gap for several reasons: First, students spend much of their time in interaction with their teachers, and teachers play a key role in determining how and what children learn (Ferguson, 2003). Second, teacher expectations have been shown to influence student outcomes (Alexander & Schofield, 2006). Third, many decisions (e.g., lesson planning, grading, tracking) are based on teachers' evaluations and predictions of students' achievement (Helmke & Schrader, 1987; Hoge & Butcher, 1984; Hoge & Coladarci, 1989). Accurate evaluations of student performance are hence highly relevant for students' learning progress. Based on the existing literature, it is unclear whether teachers have different expectations for immigrant and non-immigrant students, whether their evaluations and predictions of student performance differ across student groups, and whether these differences are caused by ethnicity or other related factors.

In the following, we first summarize research on teacher expectations for minority students, the findings of which are inconsistent. Secondly, we discuss research on teacher evaluations and predictions and how inaccurate teacher evaluations can perpetuate the achievement gap between immigrant and non-immigrant students. In the following, the terms evaluation and prediction will be used interchangeably to refer to the process of evaluating or judging a performance of a given student by the teacher. Third, we suggest that students' language background systematically influences the accuracy of teachers' evaluations.

We argue that teachers do not purposefully perpetuate the achievement gap, but that they are inadequately prepared to accurately assess the performance of immigrant students from different language backgrounds. Accordingly, we take a closer look at bilingual immigrant students and the linguistic complexity of the teaching material used.

The approach we take has been labeled 'conditional neutrality' (Ferguson, 2003). Conditional neutrality takes other relevant predictors into account when investigating ethnic bias. From this perspective, bias can be defined as the difference between teachers' assessments and a defined benchmark. Controlling for other relevant variables acknowledges potential differences between groups whose effects would otherwise be classified as bias. Teacher expectations or evaluations are thus unbiased "if they are based on legitimate observable predictors of performance" (Ferguson, 2003, p. 466).

1.1. Teacher expectations

Research has shown that teacher expectations can serve as self-fulfilling prophecies (Eckert, Dunn, Codding, Begeny, & Kleinmann, 2006; Jussim, Eccles, & Madon, 1996; Rosenthal & Jacobson, 1968) and that low teacher expectations can undermine students' academic performance (Schofield, 2006). Students from socially stigmatized groups that fare less well in school may be especially vulnerable to the negative effects of teacher expectations (Jussim & Harber, 2005). Three meta-analyses show that teachers in the United States expected more of White students than they do of African American or Hispanic students (Baron, Tom, & Cooper, 1985; Dusek & Joseph, 1983; Tenenbaum & Ruck, 2007). Auwarter and Aruguete (2008) found that teachers are likely to judge children from low-socioeconomic status as less favorably. Jussim et al. (1996) discuss the possibility that students from stigmatized groups may be more strongly affected by teachers' expectations as a result of erroneous social stereotypes. In this case, inaccurate expectations could have a self-fulfilling prophecy effect.

However, there is also evidence for a positive bias towards minority or culturally stigmatized groups of students (Jussim et al., 1996). Findings reported by Lehmann, Peek, & Gänsfuß (1997) suggest that immigrant students in Germany may even benefit from a positive grading bias. One of the reasons for these conflicting results may lie in the different methods used (Tenenbaum & Ruck, 2007).

A common methodological approach to investigating expectations involves vignettes providing varying amounts of information on students' background (Auwarter & Aruguete, 2008; Cooper, Baron, & Lowe, 1975; DeMeis & Turner, 1978). As Guttmann and Bar-Tal (1982) have shown, however, this approach has a decided tendency toward biased results. The authors conducted three studies in which they systematically varied the amount of information teachers received about students, who were to be rated on academic indicators. Results show that teachers indeed have stereotypical perceptions and expectations based on students' ethnic origin and gender. The less information was available to the teachers, however, the more pronounced the effects of stereotypical perceptions on their expectations were found to be. The results highlight the importance of investigating teacher bias in real-life situations rather than in artificial experiments (Dusek & Joseph, 1983); further, they show how closely teacher expectations and evaluations are linked. Teachers' expectations influence not only the support that teachers provide for individual students but also their evaluations of student performance.

1.2. Teacher evaluations

In their study on anti-immigrant sentiments in Danish classrooms, Wagner, Camparo, Tsenkova, and Camparo (2008) cite evidence that teachers treat and evaluate minority and majority students differently. Inaccurate teacher evaluations of students' academic abilities have been shown to hinder students' learning processes by obstructing effective teacher behaviors (Dusek & Joseph, 1983). Accurate evaluations of students' academic abilities are crucial for instructional planning and, eventually, positive student outcomes (Herman & Choi, 2008; Schrader & Helmke, 1990). Inaccurate evaluations, on the contrary, can impede the learning process: teachers can only select teaching materials on an appropriate level and offer the

Author's personal copy

support needed if they understand where their students stand. If teachers' evaluations are systematically biased, specific groups of students may benefit less from the instruction and thus be systematically disadvantaged. Such processes could result in a systematic performance gap. Previous studies of teacher prediction accuracy report an overall median correlation of .66 between teachers' predictions of student performance and actual student performance (Hoge & Coladarci, 1989). This moderate to high correlation suggests that teachers are generally quite accurate in their evaluations. Nevertheless, the range of correlations reported is wide, suggesting that moderators (e.g., student ethnicity and linguistic background) may play an important role.

In a study with 32 German classes and their teachers, Schrader and Helmke (1990) investigated which factors influence mathematics teachers' predictions of their students' performance. Each student completed a curriculum-based mathematics test, an IQ test, and a questionnaire assessing sociodemographic (e.g., nationality) and motivational variables. Teachers were asked to judge how many points students would score in the mathematics test. Results from multilevel analyses indicated that intelligence, students' mathematical self-concept, students' effort, and learning-conducive activities were significant predictors of the accuracy of teachers' predictions, but that ethnicity had no effects.

However, it is possible that no effects were found because ethnicity was operationalized in terms of nationality. Hence, the study could not distinguish between students from immigrant families who hold a German passport and students from non-immigrant families. Yet data from the microcensus show that about half of German residents with an immigration background hold a German passport (Woellert, Kröhnert, Sippel, & Klingholz, 2009). Thus, this approach fails to identify many students with an immigration background, and potentially a different language background, and may thus mask the effects of ethnicity.

1.3. *Language of instruction*

The relationship between proficiency in the language of instruction and academic outcomes has been widely discussed. Many studies show that immigrant students' lower educational outcomes are closely related to their lower reading skills (Baumert & Schümer, 2001; Lehmann et al., 1997). At the same time, there is empirical evidence that teachers have difficulty identifying students with very low levels of reading proficiency (Artelt, Stanat, Schneider, & Schiefele, 2001). Hence, students with poor reading skills are less likely to succeed in school, but their academic failure is less likely to be attributed to their reading deficiencies. As a consequence, many of these students do not receive the necessary support. In Germany, immigrant students are significantly more likely than majority students to belong to this "at-risk" group (cf. Stanat & Christensen, 2006).

It is only recently that attention has been drawn to the importance of language proficiency for academic success in subjects other than language arts (Abedi & Lord, 2001, p. 219). For example, research has shown that students tend to have more difficulty solving mathematics problems expressed in words (word problems) than in numeric format (Abedi, Lord, & Plummer, 1997). In particular, English language learners (ELL students) score lower than native speakers in mathematics proficiency tests, and the performance difference is especially pronounced on linguistically complex items (Abedi et al., 1997).

In their experiment, Abedi and Lord (2001) found that students – especially low-achieving, low socioeconomic status, and ELL students – perform better when the linguistic complexity of mathematics problems is reduced but the mathematical content kept constant. In similar vein, Wolf, Herman, et al. (2008) used a differential item functioning (DIF) approach to examine systematic bias against ELL students in standardized mathematics tests administered in three U.S. states. Wolf, Herman, et al. (2008) indeed found DIF against ELL students in some test items, meaning that an ELL student was less likely to solve the item than a non-ELL student with the same mathematical ability. The items identified featured substantially more academic English than the non-DIF items, suggesting that the abilities of ELL students may be masked by language difficulties. Drawing on these and similar findings, Abedi et al. (1997) concluded that "bilingual students keep pace with monolinguals in mechanical arithmetic but fall behind in solving word problems" (p. 5).

Mathematical word problems are often couched in academic language, and some researchers have argued that immigrant students often lack sufficient cognitive academic language proficiency (CALP), even if their basic interpersonal communication skills (BICS) are good (cf. Cummins, 2002, for a distinction of BICS and CALP). Teachers may not recognize the specific linguistic challenges that mathematical word problems pose for an immigrant student, especially if that student has good communication skills and generally good mathematical content knowledge. As a consequence, they may overestimate the student's performance (cf. Cummins, 2002).

Additionally, the linguistic complexity of word problems can vary, potentially affecting the ability of ELL students to respond (Wolf, Kao, et al., 2008). Teachers may not take such linguistic considerations into account when choosing their teaching materials. Again, the result may be that they overestimate the performance of second language learners, in particular.

1.4. *Research questions*

To summarize, student characteristics such as ethnicity or language background may influence teacher expectations, evaluations, and predictions of students' performance. However, empirical findings to date are mixed, and the precise roles played by ethnicity and language in teachers' evaluations remain unclear.

In an attempt to close this research gap, the present study investigated the accuracy of mathematics teachers' predictions of students' performance on mathematics problems, taking the students' language background into account. Because

Author's personal copy

previous research has shown that these relations should preferably be investigated in real-life settings, where teachers can be asked to evaluate their own students (Dusek & Joseph, 1983), we used data from a large-scale field assessment rather than vignettes.

In a first step, we investigated the accuracy of mathematics teachers' predictions of student performance in general. We hypothesized that mathematics teachers would *not* be more likely to overestimate the performance of bilingual immigrant students relative to the performance of non-immigrant or monolingual immigrant students.

Instead we hypothesized that teachers' accuracy in predicting immigrant students performance is related to linguistic complexity of the task. Therefore, we investigated whether teachers' accuracy in predicting immigrant students' performance differs between linguistically simple and complex mathematical problems. We hypothesized that mathematics teachers would only overestimate the performance of bilingual immigrant students relative to that of non-immigrant or monolingual immigrant students on linguistically complex problems.

2. Method

2.1. Sample

Data were collected as part of the COACTIV study (Kunter et al., 2007), which was embedded in the 2003/2004 cycle of the German national component of the Programme for International Student Assessment (PISA, Prenzel et al., 2006). PISA is an internationally standardized assessment that is administered to 15-year-olds in schools around the globe at 3-year intervals. Data for the present study come from the 2003 assessment, which investigated student achievement in three domains – mathematics, reading, and science – with a focus on mathematics literacy. As part of the German national extension to PISA 2003, the COACTIV study assessed the teachers teaching in the respective PISA classrooms. The data set thus draws on a representative sample of teachers of grade 9 mathematics in Germany, which includes teachers from all German school tracks. As a part of teacher competence, the COACTIV study also assessed teachers' diagnostic evaluations of student achievement. The project group explicitly chose mathematical problems that were given to the students and to the teachers to predict the students' performance (for teacher competences see also Blum, Neubrand, & Krauss, 2008; Krauss, Baumert, & Blum, 2008; Krauss, Brunner, et al., 2008). The analyses of the present study are based on this prior selection by the research group.

2.1.1. Participating teachers

The 305 mathematics teachers (41% female) in the present sample were on average $M = 48.50$ years old ($SD = 8.58$ years; range: 26–65 years). On average, they had been in the profession for $M = 22.36$ years ($SD = 10.19$ years; range: 3–42 years) and had been teaching the PISA classes for 0–5 years ($M = 1.89$, $SD = 1.17$). Finally, 85.3% had majored in mathematics at university.

2.1.2. Participating students

Seven students were randomly selected from each of the 305 classes taught by the teachers in our sample, and the teachers rated these students' performance on given mathematics problems that were included in the PISA test. Of the 305 teachers, 36 rated only six students, 4 teachers rated only five students, and 1 rated only four students. Hence, the final sample comprised $N = 2088$ students. Because of the real-life sample, not all teachers rated students from all three student groups (German, monolingual immigrant, bilingual immigrant). On average, students (50.5% female) were 15.25 years of age ($SD = 0.66$; range: 13–18 years). The students were very heterogeneous in terms of socioeconomic status (ISEI: $M = 50.94$, $SD = 16.2$; sample range 16–90, scale is described in the measures section). Further, 543 students (26%) reported that at least one of their parents was foreign-born (immigrant students) and 152 students (8%) spoke a language other than German at home with their families. Of the immigrant students, 149 (35%) reported that they spoke a language other than German at home.

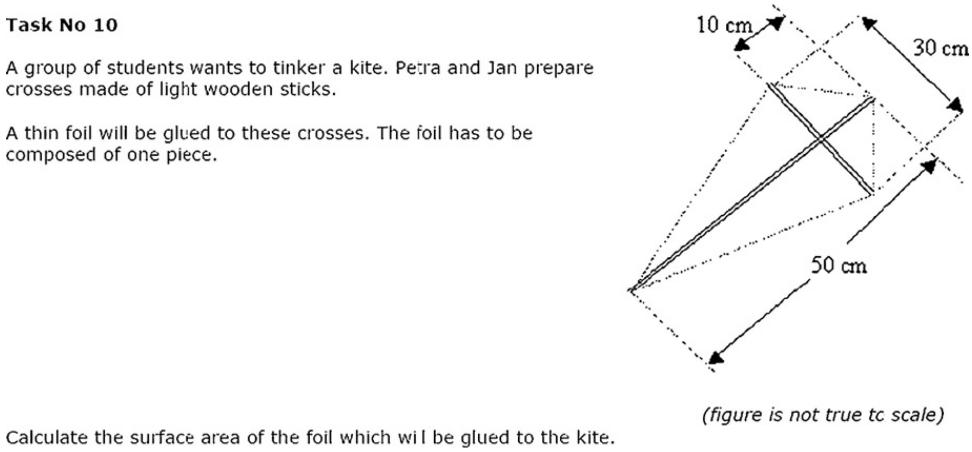
2.2. Measures

2.2.1. Independent variables

2.2.1.1. Mathematics problems. Two mathematics problems (see Figs. 1 and 2) from PISA 2003 for which teachers' predictions of students' performance and actual student performance data were available were chosen (Jordan et al., 2008). These problems displayed comparable mathematical complexity but differed in terms of linguistic complexity, making them especially suitable for investigating how linguistic complexity interacts with students' language background to influence the accuracy of teachers' evaluations.

The first problem was a geometric one that asked students to calculate the surface area of a kite. The second problem required knowledge of percentage calculation. Both problems exhibited curricular validity but differed with respect to their linguistic complexity and hence the language-related demands on the problem solver. In the *kite* problem, all relevant metrics could be read off the geometrical figure. In the *percentage* problem, the sentences had to be deconstructed and the relevant figures found and translated into the right equation. Therefore, any misunderstanding of the wording or phrasing of the problem would result in the wrong answer.

Author's personal copy



Task No 10

A group of students wants to tinker a kite. Petra and Jan prepare crosses made of light wooden sticks.

A thin foil will be glued to these crosses. The foil has to be composed of one piece.

(figure is not true tc scale)

Calculate the surface area of the foil which wi l be glued to the kite.

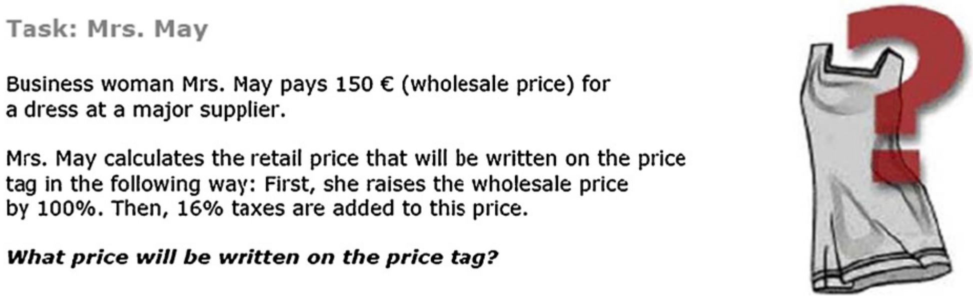
Fig. 1. Linguistically low mathematics problem (*kite* problem).

Within the German PISA assessment framework, the mathematical complexity of the problems were rated and they were classified as comparable regarding their curricular level, the problem category, the intensity of mathematical ideas implicit in the problems, and the low level of mathematical reasoning required for solving each problem. However, they differed slightly in what the PISA classification refers to as 'inner-mathematical modeling', that is the degree to which students need to translate mathematical ideas, and relate and coordinate the (mathematical) information provided by the problem (cf. Jordan et al., 2006). Whereas the *percentage* problem requires no 'inner-mathematical modeling', the *kite* problem is classified on a low level. Still, across all students, the probability of giving the correct answer did not differ between the two problems ($t(2087) = -1.02$, *n.s.*). The predictions of the teachers, however, differed significantly for the two problems ($t(1984) = -7.71$, $p > .001$). In line with the expert ratings, teacher estimates of percentage correct for all students was higher for the kite problem. Apparently, teachers rated the kite problem as mathematical more complex and hence more difficult to solve than the percentage problem.

To assess differences in the linguistic complexity of the problems in more detail, we asked 12 experts in mathematics and education to rate the degree of language proficiency needed to solve each problem on rating scales adopted from Wolf, Hermann, et al. (2008). The first rating scale tapped the extent to which a test taker has to draw on language proficiency to solve the test item correctly. The scale ranged from 0 ("no language proficiency required") to 3 ("language proficiency required to understand the relations between sentences"). For the *percentage* problem, all raters agreed that a high level of language proficiency was needed ($M = 3.00$, $SD = 0$). For the *kite* problem, language proficiency was rated to be less important ($M = 2.08$, $SD = 0.29$). The difference in ratings between the problems was statistically significant ($t(11) = -11.00$, $p < .001$). The second rating scale assessed linguistic vs. non-linguistic aspects of the test item. The scale ranged from 0 ("item consists entirely of non-linguistic features") to 3 ("item consists entirely of linguistic features"). The expert ratings indicate that the *percentage* problem consists mainly of linguistic features ($M = 2.67$, $SD = 0.49$) and the *kite* problem more of non-linguistic features ($M = 1.17$, $SD = 0.39$). The difference in ratings between the problems was again significant ($t(11) = -6.51$, $p < .001$).

In sum, the experts concluded that the problems differed substantially with regard to their linguistic complexity, linguistic features, and the demands made on students' language skills. Linguistic complexity was classified as high for the *percentage* problem and as low for the *kite* problem, and dummy coded (1 = high). In the following, we thus refer to the *percentage* problem as the linguistically high problem and to the kite problems as the linguistically low problem.

2.2.1.2. Student characteristics. Mathematical achievement, reading achievement, and socioeconomic status were included in the background model and served as control variables. Mathematical and reading achievement were operationalized by students' *z*-scores on the PISA test (Prenzel et al., 2006). *z*-Scores were used to interpret results as standardized coefficients.



Task: Mrs. May

Business woman Mrs. May pays 150 € (wholesale price) for a dress at a major supplier.

Mrs. May calculates the retail price that will be written on the price tag in the following way: First, she raises the wholesale price by 100%. Then, 16% taxes are added to this price.

What price will be written on the price tag?

Fig. 2. Linguistically high mathematics problem (*percentage* problem).

Author's personal copy

Socioeconomic status (SES) was assessed by the *International Socio-Economic Index of Occupational Status* (ISEI, Ganzeboom, De Graaf, Treimann, & De Leeuw, 1992). The ISEI is based on international data on the income and educational background of different vocations. The scale covers the theoretical range from 16 (low SES; e.g., cleaning person) to 90 (high SES; e.g., judge).

Students' immigration and language background was of special theoretical interest. Following the PISA 2003 assessment (Prenzel et al., 2006), immigration background was operationalized in terms of students' and parents' place of birth, with students who were foreign-born or who had at least one foreign-born parent being characterized as immigrant. Operationalizing immigration background via place of birth of participants and their parents instead of via nationality only has the advantage that children with immigrant background but German passport are included in the immigrant group. Information on students' language background was obtained by a question asking which language they spoke most often at home. For the present study, data on immigration background was combined with data on the language spoken at home to distinguish three groups of students: non-immigrant (German) students (80%), German-monolingual immigrant students (monolingual, 7%), and immigrant students who speak a language other than German at home (bilingual, 13%).

2.2.2. Dependent variable

The dependent variable was teachers' accuracy in predicting students' ability to solve the two mathematics problems chosen from the PISA test (Prenzel et al., 2006). Teachers were asked to predict whether each of the randomly selected students would be able to solve each problem. These predictions were then compared with the student's actual performance. Hence, the dependent variable was not students' achievement, but the accuracy of teachers' predictions of students' achievement. We chose this combined outcome variable, because it gives a direct measure of teachers' predictive accuracy. Teacher responses were coded as follows: When teachers overestimated student performance, predicting that a student who in fact failed to do so would solve the problem, responses were coded as +1. When teachers predicted student performance correctly, responses were coded as 0 (whether the student solved the problem correctly or not). Finally, when teachers underestimated student performance, predicting that a student who in fact gave the right answer would not solve the problem, responses were coded as -1. Consequently, values higher than 0 indicate overestimation, whereas values lower than 0 indicate underestimation of student performance. A value of 0 corresponds to a perfect match between the teacher's predictions and the student's actual performance. Because our hypotheses focused on teacher overestimation of student performance, we were specifically interested in values in the positive range between 0 and 1.

2.3. Analyses

2.3.1. Missing data

All students were administered a mathematics assessment including the two focal mathematics problems and a sociodemographic questionnaire. Due to the rotation design implemented in PISA 2003, reading achievement data were available for only 54% of the students. Hence, reading achievement data are *missing by design*, while all other missing values are *missing at random* (MAR). Missing data represent a potentially serious methodological problem in any study for three reasons: loss of efficiency due to reduced sample size, biased estimations due to differences between observed and non-observed data, and difficulty dealing with the data because most standard statistical packages depend on complete data matrices (Peugh & Enders, 2004). There is growing consensus – especially in the case of missing by design or MAR – that imputation of missing observations is preferable to pairwise or listwise deletion (Schafer & Graham, 2002). We therefore used a multiple imputation method to estimate missing observations. This method produces several independent data sets taking estimation errors into account and can be used even when the proportion of missing data is high (up to 50%, Graham, 2009). The auxiliary variables we used for the imputation procedure were students' characteristics (e.g., immigration and language background, gender), SES, mathematics and reading achievement, cognitive ability, and other school-related variables. Using the NORM software (version 2.03, Schafer, 2000), we generated ten data sets in which all missing data were replaced with imputed values. All subsequent data analyses were applied to these ten datasets, which were then combined according to the procedure proposed by Rubin (1987).

2.3.2. Multilevel modeling

Multilevel analyses assessed the effects of immigration and language background on the accuracy of teachers' evaluations of students' performance while controlling for students' mathematical and reading achievement and SES. In most studies conducted in the school setting, student and classroom characteristics are confounded because students are not randomly assigned to groups. In our data set, students were not randomly assigned to teachers, and each teacher rated several students: Teachers (level 3) rated the performance of each student (level 2) on two mathematics problems (level 1). We used a three-level multilevel modeling approach to handle this data structure. The teacher level (level 3) was included to account for common variance, but no predictor was entered on this level. The student level (level 2) contained individual student characteristics such as mathematics and reading achievement (*z*-scores) and SES. Level 2 also included information on immigration and language background. For the analyses, we computed two dummies. The first ("monolingual dummy") contrasted monolingual immigrant students (=1) with all other students (=0). The second ("bilingual dummy") contrasted bilingual immigrant students (=1) with all other students (=0). Both dummy variables served as predictors for the accuracy of teachers' evaluations of students' performance. Because the reference category of the two dummy variables is always coded as 0, a positive regression coefficient for the dummy variables indicates an overestimation of the monolingual immigrant or

Author's personal copy

bilingual immigrant students' performance. At level 1, linguistic complexity of the problems was entered as a dummy, with low linguistic complexity as the reference category. Random-intercept models were estimated in which the intercepts on all three levels were allowed to vary randomly but with fixed effects for all predictor variables and cross-level interactions. The model specifications for each hypothesis are presented in the Appendix. The data were analyzed using the HLM 6.0 software (Raudenbush, Bryk, Cheong, & Congdon, 2004). Estimation problems were not encountered. The method of estimation applied for all models was full maximum likelihood, using an empirical Bayes algorithm. Hypotheses about fixed effects were tested. Because the continuous nature of the data is debatable, we repeated our analyses with a binary outcome variable (student overestimated vs. student underestimated) using logistic regression (see Raudenbush & Bryk, 2002). Results were nearly identical. To avoid redundancy and to simplify the presentation and interpretation of results, we therefore present only the analyses using over- and underestimation simultaneously as the outcome variable.

Hypothesis 1 states that mathematics teachers do *not* overestimate the performance of bilingual immigrant students relative to that of monolingual immigrant or non-immigrant students in general. To test hypothesis 1, we examined the main effect of language and immigration background on the accuracy of teachers' predictions.

Hypothesis 2 tests the interaction and states that mathematics teachers overestimate the performance of bilingual immigrant students on linguistically high problems. To test hypothesis 2, we examined the cross-level interaction between immigration and language background (level 2) and the complexity of the task (level 1).

3. Results

3.1. Descriptive results

Table 1 shows the percentages of students who solved the problems correctly by linguistic complexity of the problem and student group (non-immigrant, monolingual immigrant, bilingual immigrant). In all cases, less than one-third of the students solved the problems correctly. For the linguistically low problem, we found no performance differences between the student groups ($\chi^2(2, N = 2022) = 3.13$, ns). For the linguistically high problem, we found a significant difference in student performance ($\chi^2(2, N = 2047) = 20.45$, $p < .001$), with the group of bilingual immigrant students showing a significantly lower percentage correct than the other two groups.

Table 1 also reports teachers' predictions of student performance (percentage correct for all students). In all cases, teachers estimated that at least half of the students would solve the problems correctly. Teacher predictions were slightly higher for the linguistically high problem than for the linguistically low one. However, teacher predictions did not differ significantly across the three student groups—for either the linguistically low problem ($\chi^2(2, N = 2083) = 2.89$, ns) or the linguistically high problem ($\chi^2(2, N = 1989) = 0.43$, ns). In sum, Table 1 shows that bilingual immigrant students performed significantly worse on the linguistically high problem than the other two groups. However, this difference was not reflected in the teachers' predictions.

Comparison of students' actual performance and teachers' predictions yielded the dependent variable shown in the last column of Table 1. Positive values indicate that teachers overestimated student performance. Teachers' predictions differed significantly from 0 for both the linguistically low problem ($t(2082) = 25.49$, $p < .001$) and the linguistically high problem ($t(1988) = 28.76$, $p < .001$), indicating that teachers generally tended to overestimate the performance of their students.

The accuracy of teachers' evaluations of student performance on the two problems was positively correlated ($r = .24$, $p < .001$, $N = 1985$); teachers who overestimated their students' performance on one problem were more likely to overestimate performance on the other problem as well.

To test whether teachers overestimated the different student groups equally on both problems, we computed a 2 (linguistic complexity: high vs. low) × 3 (student groups: non-immigrant, monolingual immigrant, bilingual immigrant) ANOVA. Overall, there was no difference in the degree to which the teachers overestimated the three student groups, $F(2, 1982) = 1.54$, ns, but they did overestimate students' performance on the linguistically high problem more strongly than

Table 1
Percentage of students who solved the mathematics problems correctly (student performance) and teacher estimates in terms of percentage correct for all students (teacher predictions) and accuracy of teachers' predictions of students' performance by student group and linguistic complexity of the problem.

Group	Student performance: percentage correct	Teacher predictions: teacher estimates of percentage correct for all students	Mean accurarcy of teachers' predictions of students' performance (*SD*)
Linguistically low problem			
Non-immigrant students	27%	58%	.32 (.58)
Monolingual immigrant students	23%	57%	.34 (.58)
Bilingual immigrant students	21%	51%	.31 (.56)
Linguistically high problem			
Non-immigrant students	29%	66%	.37 (.62)
Monolingual immigrant students	23%	68%	.44 (.60)
Bilingual immigrant students	12%	66%	.53 (.54)

Author's personal copy

their performance on the linguistically low problem, $F(1, 1982) = 21.78$, $p < .001$. The interaction term was also significant, $F(2, 1982) = 4.21$, $p < .05$. As post hoc tests showed, the student groups did not differ for the linguistically low problem, $F(2, 1982) = 0.56$, ns, but for the linguistically high problem, with teachers overestimating bilingual immigrant students' performance on this problem significantly more strongly than they did the performance of the other two groups, $F(2, 1982) = 4.36$, $p < .01$. These analyses provide first support for our hypotheses, but do not take the multilevel structure of our data into consideration. Hence, in a next step, we used multilevel modeling, including control variables to the satisfy requirements of a 'conditional neutrality' approach.

As mentioned above, a 'conditional neutrality' approach to studying bias takes into account additional predictors that are relevant to the outcome (Ferguson, 2003). According to this approach, teachers should expect the same performances for students of different origins only if the students are comparable on relevant variables. If students are not comparable, the relevant variables should be controlled for. Hence, our analyses included variables that are related to performance on the mathematics problems and that differ across the three student groups.

Performance on the mathematics problems was significantly correlated with mathematical achievement, reading achievement, and SES (for the linguistically low problem: .34, .23, .12; for the linguistically high problem: .38, .32, .18, respectively; all correlations $p < .001$). The differences between the correlations for the linguistically low and the linguistically high problem were significant only for reading achievement, $t(1067) = -2.48$, $p < .01$, and SES, $t(1887) = -2.14$, $p < .05$, and not for mathematical achievement ($t < 2$).

The three student groups differed significantly on all three variables (mathematical achievement: $F(2, 2067) = 36.49$, $p < .001$; reading achievement: $F(2, 1119) = 34.51$, $p < .001$; SES: $F(2, 1980) = 42.02$, $p < .001$).

Non-immigrant students had the highest SES ($M = 52.46$, $SE_M = 0.40$), followed by monolingual immigrant students ($M = 46.84$, $SE_M = 1.00$), and bilingual immigrant students ($M = 40.92$, $SE_M = -1.46$). Post hoc tests showed that all differences were significant. Effect sizes as measured with Cohen's d for the comparisons are as follows: $d = 0.35$ (non-immigrant vs. monolingual immigrant students), $d = 0.71$ (non-immigrant vs. monolingual immigrant students), and $d = 0.36$ (monolingual vs. bilingual immigrant students).

The pattern of results for mathematical achievement was the same: non-immigrant students had the highest scores ($M = 0.01$, $SE_M = 0.02$), followed by monolingual immigrant students ($M = -0.26$, $SE_M = 0.07$), and bilingual immigrant students ($M = 0.67$, $SE_M = 0.09$). Again, post hoc tests showed that all differences were significant (non-immigrant vs. monolingual immigrant students: $d = 0.27$, non-immigrant vs. monolingual immigrant students: $d = 0.67$, and monolingual vs. bilingual immigrant students: $d = 0.38$).

Post hoc tests for reading achievement revealed a slightly different pattern. The best results in the reading test were achieved by non-immigrant ($M = -0.02$, $SE_M = 0.03$) and monolingual immigrant students ($M = -0.20$, $SE_M = 0.09$). The difference between these groups was not significant. Bilingual immigrant students performed significantly worse than these two groups ($M = -1.00$, $SE_M = 0.13$, non-immigrant vs. bilingual immigrant students: $d = 0.91$, monolingual vs. bilingual immigrant students: $d = 0.72$). In the multilevel analyses reported below, we controlled for all three variables: SES, mathematical achievement, and reading achievement.

3.2. Results from multilevel models

3.2.1. Hypothesis 1

Results for the first hypothesis are shown in the first column (Model 1) of Table 2. In Model 1, we tested whether immigration and language background influenced the accuracy of teachers' predictions. More specifically, we compared prediction accuracy across the three student groups by regressing teachers' prediction accuracy on language and immigration background, while controlling for mathematical achievement, reading achievement, and SES.

As indicated by the positive and significant intercept (see first row of Model 1), teachers generally overestimated the performance of their students. However, as predicted, there was no significant difference between overestimation of the performance of non-immigrant and immigrant students, and both group contrasts (β_{04k} and β_{05k}, ns) were not significant.

None of the three background variables (SES, mathematical achievement, reading achievement) were significant. At first sight, this result seems incongruent with the influence of the three variables on students' actual performance. However, the three background variables were only related to students' performance (percentage correct) and not to teachers' prediction accuracy.

Hence, the results do support the hypothesis that the accuracy of teachers' predictions do not differ between students with and without immigrant background in general. However, in the next step, we tested whether this pattern persisted when we controlled for the linguistic complexity of the problem. Hence, in hypothesis 2, we tested whether the interaction between students' language and immigration background and the linguistic demands of the problem affected the accuracy of teachers' evaluations.

3.2.2. Hypothesis 2

Results for the second hypothesis are presented in the second column (Model 2) of Table 2. In Model 2, we tested whether teachers overestimated the performance of bilingual immigrant students only on linguistically high problems. In addition to the variables entered in Model 1, Model 2 included the linguistic complexity of the problem (level 1) and the interaction between linguistic complexity and students' immigration and language background (cross-level interaction). As in Model 1,

Author's personal copy

Table 2
Results from multilevel analyses predicting teachers' evaluation accuracy.

Parameter	Model 1		Model 2	
	Estimate	*SE*	Estimate	*SE*
Level 1 predictors				
Intercept (π_{0jk})	.351[*]	.014	.330[*]	.017
Linguistic complexity (Π_{1jk})			.043[*]	.018
Level 2 predictors				
Student group				
Monolingual immigrant students (β_{04k})	.035	.033	−.013	.043
Bilingual immigrant students (β_{05k})	.051	.041	−.035	.053
Cross-level interactions				
Monolingual immigrants × complexity			.096	.057
Bilinguals immigrants × complexity			.172[*]	.064
Background model				
SES (β_{01k})	−.001	.001	−.001	.001
Mathematics achievement (β_{02k})	.01	.013	.010	.013
Reading achievement (β_{03k})	.019	.014	.019	.014
Variance estimates				
Level-1 variance	.267		.267	
Level-2 variance	.064		.065	
Level-3 variance	.020		.020	

Note: The level-1 intercept represents the average overestimation for non-immigrant students. The problem variable is dummy coded (0 = low linguistic complexity); SES, mathematics, and reading achievement are grand mean centered. The monolingual dummy is coded 1 for monolingual immigrant students; the bilingual dummy is coded 1 for bilingual immigrant students.

[*] $p < .05$.

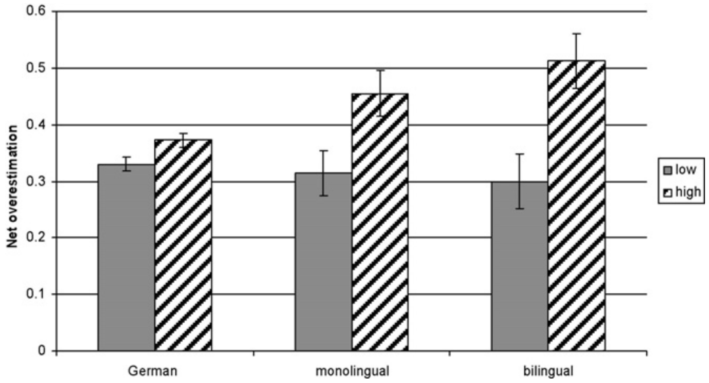


Fig. 3. Net overestimation of teachers' evaluations of students' performance on the two mathematics problems by student group. Low: linguistically low mathematics problem (*kite* problem); High: linguistically high mathematics problem (*percentage* problem). Error bars indicate standard error. The means for the monolingual immigrant and bilingual immigrant students were calculated by adding the regression coefficients of the respective dummy to the coefficients of the non-immigrant students (intercept) and the linguistically high problem.

results show that teachers generally overestimated the performance of their students. Model 2 further shows a significant main effect of the linguistic complexity of the problem (π_{1jk}): Teachers were more likely to overestimate students' performance on the linguistically high problem. Most importantly, and confirming our hypothesis, a significant cross-level interaction between immigration and language background and linguistic complexity of the problem was observed. Fig. 3 illustrates this interaction by presenting the means of each group: For the linguistically high problem, overestimation was least pronounced for non-immigrant students ($M = 0.373$, $SE_M = 0.02$), followed by monolingual immigrant students ($M = 0.469$, $SE_M = 0.15$), and most pronounced for bilingual immigrant students ($M = 0.545$, $SE_M = 0.06$). For the linguistically low problem, in contrast, there were no significant differences between the three student groups (β_{04k} and β_{05k}, ns). As in Model 1, none of the background variables were significant. In other words, mathematical achievement, reading achievement, and SES did not influence the accuracy of teachers' predictions. In sum, on the linguistically high problem, teachers overestimated the performance of bilingual immigrant students relative to that of non-immigrant students.

4. Discussion

4.1. Summary

Do teachers evaluate the performance of immigrant students less accurately than that of non-immigrant students? To address this question, we analyzed mathematics teachers' predictions of their students' performance on two mathematics

Author's personal copy

problems and compared these evaluations with the students' actual performance. Data were provided by the COACTIV study, which was embedded in the 2003 cycle of the PISA assessment in Germany.

In multilevel analyses, we tested the hypothesis that teachers systematically overestimate the performance of immigrant students relative to that of non-immigrant students. Our findings suggest that it is not ethnic background per se that affects the accuracy of teachers' predictions, but the interaction of students' ethnic and language background with the linguistic demands of the mathematics problem to be solved. Results showed that teachers tended to overestimate the performance of all students, as reported in previous studies (Anders, Kunter, Brunner, Krauss, & Baumert, 2010). Overall, we found no specific effect of immigration or language background. However, once we considered the linguistic features of the problems, we observed an interaction between students' immigration and language background and the linguistic complexity of the task. Results showed that teachers were significantly more likely to overestimate the performance of bilingual immigrant students on a linguistically high problem, even when we controlled for mathematical achievement, reading achievement, and SES. These results are consistent with the idea that teachers are not always aware of bilingual students' language difficulties or of the linguistic demands that specific mathematics problems may pose.

4.2. Limitations

Unfortunately, our data do not allow us to determine whether teachers took the differing linguistic demands of the problems and the language background of the students into account when predicting students' performance. Hence, we can only speculate that teachers overestimated the performance of bilingual immigrant students on the linguistically high problem because they did not recognize the language difficulties involved. Based on their study on the accuracy of teacher assessments of second-language students at risk for reading disability, Limbos and Geva (2001) concluded that teachers may "inappropriately use oral language proficiency as their gauge for the child's overall academic performance" (p. 149) As has also been suggested by other researchers, good oral language skills of bilingual students may mask potential language problems in the academic domain (Cummins, 2002).

Due to small group sample size, our data also do not allow us to compare the accuracy of teacher predictions for students from different countries of origin. However, it is possible that teachers predictions of the performance of students with an European background differ from those for students with a non-European background. Further research should investigate this more closely. The data did also not provide equal distribution of teacher evaluations for each of the three student groups, and some teachers evaluated only German or monolingual immigrant students. While unavoidable when using real-life data sets, this procedure might affect teachers' reference categories and thereby influencing their predictions for the individual student. However, because teachers were not asked to evaluate their students' performance relative to that of other students, their predictions should be independent.

Another limitation of the present study is that we focused on just two mathematics problems from different mathematical fields. However, both problems exhibited curricular validity and pertained to mathematical fields that the students had covered in school. Although for the students the percent correct of the problems did not differ, teacher estimates of percent correct were higher for the linguistically high problem. That is, teachers predicted that students rather solve the linguistically high than the linguistically low problem. It is conceivable that teachers found the linguistically high problem to be mathematical less complex and focused only on mathematical reasoning instead of taking linguistic complexity into account when making their prediction This interpretation is supported by the PISA expert ratings that classified the linguistically high problem as requiring less 'inner-mathematical' reasoning. To be able to generalize the results and disentangle the effects of mathematical and linguistic complexity, future analyses should be based on a broader range of items. Disentangling the effects of mathematical and non-mathematical content (e.g., language complexity) can help researchers to gain a better understanding of differences in the performance of specific student groups. Findings showing that performance differences are not solely attributable to differences in student ability, but that certain item characteristics discriminate against specific students groups, would have major implications for research and practice.

Based on their differential item functioning (DIF) analyses of items from standardized mathematics tests administered in three U.S. states, Wolf, Hermann, et al. (2008) have concluded that the language demands of test items is one of the factors contributing to the achievement gap between language minority and majority students. To our knowledge, the performance of language minority and majority students on standardized tests in Germany has not yet been compared. A first step was taken by PISA 2003. Items from the PISA test suggest that problems from some areas of the natural sciences contain more linguistic features than do problems from other areas (e.g. problems with graphs). However, first analyses do not show that immigrant students perform systematically worse on items with more linguistic features (Ramm, Prenzel, Heidemeier, & Walter, 2004). Yet, more systematic analyses are necessary to come to a final conclusion.

We have argued that the problems examined in the present study differed with respect to their linguistic complexity; however, we did not assess students' or teachers' perceptions of linguistic complexity. Instead, we obtained ratings of linguistic features from experts in mathematics and education. Although it is possible that student and teacher subjective appraisals differ from our expert ratings, previous research indicates that linguistic features influence the difficulty of mathematics problems. It has been shown that word problems are more difficult to solve than are mathematics problems presented in numeric format, and that the percentage correct is higher when problems contain less academic language (Abedi & Lord, 2001). Bilingual students, in particular, seem to perform better when the linguistic complexity of mathematics

Author's personal copy

problems is reduced (Abedi & Lord, 2001). These results support our finding that student reading literacy affects performance on mathematical word problems. We thus controlled for student reading achievement in all multilevel analyses predicting teachers' prediction accuracy.

4.3. Strengths

A major advantage of the present study is that it examined the accuracy of teacher predictions in a real-life (classroom) situation in which teachers knew their students well. Previous research has shown that teachers who have little knowledge about the students they are asked to evaluate rely on biasing information (Guttmann & Bar-Tal, 1982). Hence, investigating bias in a naturalistic setting strengthens the ecological validity of the findings. Another advantage is that the assessment was embedded in the national component of the PISA study and could therefore draw on a large and representative students sample with high response rates.

The approach we have taken to investigate *general* bias in teachers' evaluations has been termed 'conditional neutrality' (Ferguson, 2003). According to this approach, teachers should expect the same performances for students of different origins *only* if the students' competences and test scores are comparable. Hence, neutrality is conditioned on observable and measurable criteria. We were able to control for mathematical achievement, reading achievement, and SES while testing for a general overestimation of immigration and language background. We operationalized immigration background in terms of the parents' and child's place of birth and included information on the language spoken at home. We were thus able to differentiate between monolingual immigrant students (who have at least one foreign-born parent but speak German at home) and bilingual immigrant students. But our analyses moved beyond testing for a general overestimation of immigrant students. Our study shifts the focus from investigating general bias and instead examines the specific student–task interactions that affect teachers' predictive accuracy. Future studies should continue to investigate such interplays, how it affects teachers, and how teachers can be sensitized for such interactions.

4.4. Implications

The findings presented have several implications. First and foremost, our results do not support the claim made elsewhere that teachers have lower expectations for immigrant students. On the contrary, teachers even overestimated their performance. While this may sound as good news, overestimation might impede effective teaching just as much as underestimation as we will discuss below.

Secondly, the results again highlight the importance of proficiency in the language of instruction. Most studies to date have examined the direct role of proficiency in the language of instruction for academic outcomes. Clearly, students who do not speak the language of instruction will not be able to follow, to participate in, or to benefit from lessons.

Going beyond this, our study shows that language proficiency can also influence academic success more indirectly: namely via its impact on teachers' evaluations. Our results suggest that mathematics teachers ignore the interplay between linguistic complexity of mathematical problems and linguistic background, leading to less accurate predications of the performance of bilingual immigrant students. However, predicting student performance is an integral part of teachers' daily professional lives (Demaray & Elliot, 1998; Hoge & Butcher, 1984) and the ability to make accurate predictions is an important aspect of teaching skills. It enables teachers to adapt their lessons to students' needs, to choose appropriate teaching materials and assessment tools, and to provide effective learning opportunities (cf. Anders et al., 2010; Edelenbos & Kubanek-German, 2004).

Inaccurate evaluations, in turn, interfere with all of the above-mentioned processes and can thus impede students' learning progress (for mathematics, cf. Anders et al., 2010). Accurately evaluating what students understand and where their deficits are during the learning process can be seen as the basis for adaptive teaching (Baumert & Kunter, 2006).

In one of the few studies investigating the effects of judgment accuracy of teachers on student achievement, Helmke and Schrader (1987) showed that high judgment accuracy combined with – and only if combined with – appropriate instructional techniques were particularly favorable for classroom growth of achievement. The authors argue that "diagnostic sensitivity, therefore, can be regarded as a necessary precondition for the successful use of structuring cues." (p. 96) At the same time, high judgment accuracy was negatively related to growth of achievement when individual support was not provided by the teacher. While these results applied for all students, they might be especially relevant when considering implications of our findings for bilingual immigrant students. If quality of instruction is influence by the diagnostic sensitivity of teachers (Helmke & Schrader, 1987, p. 97), it is problematic when the diagnostic sensitivity systematically differs for different students groups. Although our data do not allow conclusions about the individual support that teachers supply for their students, an overestimation of bilingual students' performance could result in less support for that particular group. More research is still necessary to investigate implications of diagnostic accuracy on the learning progress of students.

International large-scale assessments have documented an academic achievement gap between majority and minority students, especially in Germany (Stanat & Christensen, 2006). Educational and social scientists have investigated several potential causes of this ethnic achievement gap, such as differences in allocation of resources (including financial, cultural, and social capital), and differences in aspirations related to education; however, the role of teachers and how they evaluate students of different origins remains an important issue.

Author's personal copy

Previous research has focused on potential bias in teachers' expectations, assessments, beliefs, and behavior. In this article, however, we have argued that teachers may have difficulty to accurately evaluate, and thus appropriately support, minority students' performance because they fail to recognize these students' language difficulties. Another implication we can draw is thus that teachers need to be trained to assess the linguistic complexity of their teaching materials and to recognize their students' language problems (especially where bilingual students are concerned). As our study shows, this does not apply solely to teachers of language arts. Accurate evaluations of what students know and where they need special assistance are vital for improving the quality of instruction. Chang (2008) has argued that improving the quality of classroom practice will lead to better cognitive outcomes, especially for language minority students. Her research has revealed that teacher-directed whole-class activities pose problems for language minority students. One reason for this finding may be that such activities do not give teachers the opportunity to obtain feedback on the accuracy of their evaluations and to learn more about their students' individual problems (e.g., language or comprehension problems). If teachers fail to identify students' learning problems, however, those students will fall behind. In the case of students with immigration background, this is clearly already happening—not only in Germany.

Acknowledgement

The authors thank Susannah Goss for editorial assistance.

Appendix A. Appendix

Level	Model	Description
Equations for model 1		
Level 1:	$Y_{ijk} = \pi_{0jk} + e_{ijk}$	(Level 1)
Level 2:	$\pi_{0jk} = \beta_{00k} + \beta_{01k} X_{1jk} + \beta_{02k} X_{2jk} + \beta_{03k} X_{3jk} + \beta_{04k} X_{4jk} + \beta_{05k} X_{5jk} + r_{0jk}$	(Model for level-1 intercept parameter plus random component)
Level 3:	$\beta_{00k} = \gamma_{000} + u_{00k}$	(Model for level-2 intercept parameter plus random component)
	$\beta_{01k} = \gamma_{010}$	(Model for level-2 slope parameter)
	$\beta_{02k} = \gamma_{020}$	(Model for level-2 slope parameter)
	$\beta_{03k} = \gamma_{030}$	(Model for level-2 slope parameter)
	$\beta_{04k} = \gamma_{040}$	(Model for level-2 slope parameter)
	$\beta_{05k} = \gamma_{050}$	(Model for level-2 slope parameter)
Equations for model 2		
Level 1:	$Y_{ijk} = \pi_{0jk} + \pi_{1jk} E_{1jk} + e_{ijk}$	(Level 1)
Level 2:	$\pi_{0jk} = \beta_{00k} + \beta_{01k} X_{1jk} + \beta_{02k} X_{2jk} + \beta_{03k} X_{3jk} + \beta_{04k} X_{4jk} + \beta_{05k} X_{5jk} + r_{0jk}$	(Model for level-1 intercept parameter plus random component)
	$\pi_{1jk} = \beta_{10k} + \beta_{11k} X_{4jk} + \beta_{12k} X_{5jk} + r_{0jk}$	(Model for level-1 slope parameter plus random component)
Level 3:	$\beta_{00k} = \gamma_{000} + u_{00k}$	(Model for level-2 slope parameter plus random component)
	$\beta_{01k} = \gamma_{010}$	(Model for level-2 slope parameter)
	$\beta_{02k} = \gamma_{020}$	(Model for level-2 slope parameter)
	$\beta_{03k} = \gamma_{030}$	(Model for level-2 slope parameter)
	$\beta_{04k} = \gamma_{040}$	(Model for level-2 slope parameter)
	$\beta_{05k} = \gamma_{050}$	(Model for level-2 slope parameter)
	$\beta_{10k} = \gamma_{100}$	(Model for cross-level interaction)
	$\beta_{11k} = \gamma_{110}$	(Model for cross-level interaction)
	$\beta_{12k} = \gamma_{120}$	(Model for cross-level interaction)

Criterion and predictor variables: Y_{ijk}: criterion; teachers' overestimation of students' performance for word problem *i* and student *j* in classroom *k*. Π_{1jk}: Level-1 predictor; linguistic complexity of problem for student *j* in classroom *k* (dummy coded: 0 = linguistically low, 1 = linguistically high). X_{1jk}: Level-2 predictor; socioeconomic background of student *j* in classroom *k*; X_{2jk}: Level-2 predictor; mathematical achievement of student *j* in classroom *k*. X_{3jk}: Level-2 predictor; reading achievement of student *j* in classroom *k*. X_{4jk}: Level-2 predictor; immigration background of student *j* in classroom *k* (monolingual dummy; 1 = monolingual immigrant students). X_{5jk}: Level-2 predictor; language background of student *j* in classroom *k* (bilingual dummy; 1 = bilingual immigrant students).

References

Abedi, J., & Lord, C. (2001). The language factor in mathematics tests. *Applied Measurement in Education, 14*, 219–234.

Abedi, J., Lord, C., & Plummer, J. R. (1997). *Final report of language background as a variable in NAEP mathematics performance (CSE Technical Report No. 429)*. Los Angeles, CA: University of California, Los Angeles, Center for the Study of Evaluation, National Center for Research on Evaluation, Standards, and Student Testing (CRESST).

Alexander, K. L., & Schofield, J. W. (2006). Expectancy effects: How teachers' expectations influence student achievement. In J. W. Schofield (Ed.), *Migration background, minority-group membership, and academic achievement research. Evidence from social, educational, and developmental psychology*. Berlin: Social Science Research Center.

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

Author's personal copy

Artelt, C., Stanat, P., Schneider, W., & Schiefele, U. (2001). Lesekompetenz: Testkonzeption und Ergebnisse [Reading literarcy: Test conception and results]. In J. Baumert, E. Klieme, M. Neubrand, M. Prenzel, U. Schiefele, W. Schneider, P. Stanat, K.-J. Tillmann, & M. Weiß (Eds.), *PISA 2000. Basiskompetenzen von Schülerinnen und Schülern im internationalen Vergleich* (pp. 69–137). Opladen: Leske+Buderich.

Auwarter, A. E., & Aruguete, M. S. (2008). Effects of student gender and socioeconomic status on teacher perceptions. *Journal of Educational Research, 101*, 243–246.

Baron, R. M., Tom, D. Y. H., & Cooper, H. (1985). Social class, race and teacher expectations. In J. B. Dusek (Ed.), *Teacher expectancies* (pp. 251–270). Hillsdale, NJ: Erlbaum.

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

Baumert, J., & Schümer, G. (2001). Familiäre Lebensverhältnisse, Bildungsbeteiligung und Kompetenzerwerb [Family background, educational participation, and competency acquisition]. In J. Baumert, E. Klieme, M. Neubrand, M. Prenzel, U. Schiefele, W. Schneider, P. Stanat, K.-J. Tillmann, & M. Weiß (Eds.), *PISA 2000. Basiskompetenzen von Schülerinnen und Schülern im internationalen Vergleich* (pp. 323–407). Opladen: Leske+Budrich.

Blum, W., Neubrand, M., & Krauss, S. (2008). *Zusammenhänge des Professionswissens mit Lehrermerkmalen, mit Unterrichtsqualität und mit dem Leistungszuwachs der SchülerInnen [Relation of professional knowledge with teacher characteristics, with quality of instruction and with student learning].* Beiträge zum Mathematikunterricht 2008. Vorträge auf der 42. Tagung für Didaktik der Mathematik vom 13.3. bis 18.3.2007 in Budapest, Münster: Martin Stein Verlag. pp. 157–160.

Chang, M. (2008). Teacher instructional practices and language minority students: A longitudinal model. *The Journal of Educational Research, 102*, 83–97.

Cooper, H. M., Baron, R. M., & Lowe, C. A. (1975). The importance of race and social class information in the formation of expectancies about academic performance. *Journal of Educational Psychology, 67*, 312–319.

Cummins, J. (2002). BICS and CALP. In M. Byram (Ed.), *Encyclopedia of language and teaching* (pp. 76–79). London: Routledge.

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

DeMeis, D. K., & Turner, R. R. (1978). Effects of students' race, physical attractiveness, and dialect on teachers' evaluations. *Contemporary Educational Psychology, 31*, 77–86.

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

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

Edelenbos, P., & Kubanek-German, A. (2004). Teacher assessment: The concept of 'diagnostic competence'. *Language Testing, 21*, 259–283.

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

Ganzeboom, H. B. G., De Graaf, P. M., Treimann, D., & De Leeuw, J. (1992). A standard international socioeconomic index of occupational status. *Social Science Research, 21*, 1–56.

Graham, J. W. (2009). Missing data analysis: Making it work in the real world. *Annual Review of Psychology, 60*, 549–576.

Guttmann, J., & Bar-Tal, D. (1982). Stereotypic perceptions of teachers. *American Educational Research Journal, 19*, 519–528.

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

Herman, J. L., & Choi, K. (2008). *Formative assessment and the improvement of middle school science learning: The role of teacher accuracy (CSE Technical Report No. 740).* Los Angeles, CA: University of California Los Angeles Center for the Study of Evaluation National Center for Research on Evaluation Standards and Student Testing(CRESST).

Hoge, R. D., & Butcher, R. (1984). Analysis of teacher judgments of pupil achievement levels. *Journal of Educational Psychology, 76*, 777–781.

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

Jordan, A., Ross, N., Krauss, S., Baumert, J., Blum, W., Neubrand, M., et al. (2006). *Klassifikationsschema für Mathematikaufgaben: Dokumentation der Aufgabenklassifikation im COACTIV-Projekt [Classification scheme for mathematical problems: Documentation of the problem classification in the COACTIV project].* Materialien aus der Bildungsforschung (vol. 81), Berlin: Max-Planck-Institut für Bildungsforschung.

Jordan, A., Krauss, S., Löwen, K., Kunter, M., Baumert, J., Blum, W., et al. (2008). Aufgaben im COACTIV-Projekt: Zeugnisse des kognitiven Aktivierungspotentials im deutschen Mathematikunterricht [Mathematical problems in the COACTIV project: Cognitive activation in German mathematical instruction]. *Journal für Mathematikdidaktik, 29*(2), 83–107.

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

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.

Krauss, S., Baumert, J., & Blum, W. (2008). Secondary mathematics teachers' pedagogical content knowledge and content knowledge: Validation of the COACTIV constructs. *The International Journal on Mathematics Education, 40*, 873–892.

Krauss, S., Brunner, M., Kunter, M., Baumert, J., Blum, W., Neubrand, M., et al. (2008). Pedagogical content knowledge and content knowledge of secondary mathematics teachers. *Journal of Educational Psychology, 100*, 716–725.

Kunter, M., Klusmann, U., Dubberke, T., Baumert, J., Blum, W., & Brunner, M. (2007). Linking aspects of teacher competence to their instruction. Results from the COACTIV project. In M. Prenzel (Ed.), *Studies on the educational quality of schools. The final report on the DFG Priority Programme* (pp. 32–53). Muenster: Waxmann.

Lehmann, R. H., Peek, R., & Gänsfuß, R. (1997). *Aspekte der Lernausgangslage von Schülerinnen und Schülern der fünften Klassen an Hamburger Schulen [Aspects of competence of 5th grade students in Hamburg schools].* Hamburg: Behörde für Schule, Jugend und Berufsbildung, Amt für Schule.

Limbos, M., & Geva, E. (2001). Accuracy of teacher assessments of second-language students at risk for reading disability. *Journal of Learning Disabilities, 34*, 136–151.

Organisation for Economic Cooperation and Development [OECD]. (2004). *Messages from the Programme for International Student Assessment*. Paris: OECD.

Peugh, J. L., & Enders, C. K. (2004). Missing data in educational research: A review of reporting practices and suggestions for improvement. *Review of Educational Research, 74*, 525–556.

Phillips, D. C., & Callan, E. (2001). Philosophy, multiculturalism and education [Special Issue]. *International Journal of Educational Research, 35*, 249–349.

Prenzel, M., Baumert, J., Blum, W., Lehmann, R., Leutner, D., Neubrand, M., et al. (2006). *PISA 2003. Untersuchungen zur Kompetenzentwicklung im Verlauf eines Schuljahres [PISA 2003 Studies on competence development during a school year].* Münster: Waxmann.

Ramm, G., Prenzel, M., Heidemeier, H., & Walter, O. (2004). Soziokulturelle Herkunft: Migration [Sociocultural background: Migration]. In M. Prenzel, J. Baumert, W. Blum, R. Lehmann, D. Leutner, M. Neubrand, R. Pekrun, H.-G. Rolff, J. Rost, & U. Schiefele (Eds.), *PISA 2003: Der Bildungsstand der Jugendlichen in Deutschland - Ergebnisse des zweiten internationalen Vergleiches*. Münster: Waxmann.

Raudenbush, S. W., & Bryk, A. S. (2002). *Hierarchical linear models* (2nd ed.). Thousand Oaks: Sage.

Raudenbush, S. W., Bryk, A. S., Cheong, Y. F., & Congdon, R. T. (2004). *HLM 6; Hierarchical linear and nonlinear modeling.* Lincolnwood, IL: Scientific Software International.

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

Rubin, D. B. (1987). *Multiple imputation for nonresponse in surveys*. New York: Wiley.

Schafer, J. L. (2000). *NORM 2.03 for Windows 95/98/NT [Computer program]* Retrieved 02/01/2008 from http://www.stat.psu.edu/~jls.

Schafer, J. L., & Graham, J. W. (2002). Missing data: Our view of the state of the art. *Psychological Methods, 7*, 147–177.

Schofield, J. W. (Ed.). (2006). *Migration background, minority-group membership, and academic achievement research. Evidence from social, educational, and developmental psychology*. Berlin: Social Science Research Center.

Schrader, F.-W., & Helmke, A. (1990). Lassen sich Lehrer bei der Leistungsbeurteilung von sachfremden Gesichtspunkten leiten? Eine Untersuchung zu Determinanten diagnostischer Lehrerurteile [Are teachers influenced by extrinsic factors when evaluating scholastic performance? A study on the determinants of teachers' judgments]. *Zeitschrift für Entwicklungspsychologie und Pädagogische Psychologie, 22*, 312–324.

Author's personal copy

Stanat, P., & Christensen, G. (2006). *Where immigrant students succeed. A comparative review of performance and engagement in PISA 2003*. Paris: OECD.

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

Wagner, J. T., Camparo, L. B., Tsenkova, V., & Camparo, J. C. (2008). Do anti-immigrant sentiments track into Danish classrooms? Ethnicity, ethnicity salience, and bias in children's peer preferences. *International Journal of Educational Research, 47*, 312–322.

Woellert, F., Kröhnert, S., Sippel, L., & Klingholz, R. (2009). *Ungenutze Potentiale. Zur Lage der Integration in Deutschland [Unutilized Potentials on the current state of integration in Germany]*. Berlin: Berlin Institute for Population and Development.

Wolf, M. K., Herman, J. L., Kim, J., Abedi, J., Leon, S., Griffin, N., et al. (2008). *Providing validity evidence to improve the assessment of English language learners (CSE Technical Report No. 738)*. Los Angeles, CA: University of California, Los Angeles, Center for the Study of Evaluation, National Center for Research on Evaluation, Standards, and Student Testing (CRESST).

Wolf, M. K., Kao, J., Herman, J. L., Bachman, L. F., Bailey, A., Bachman, P. L., et al. (2008). *Issues in assessing English language learners: English language proficiency measures and accommodation uses. Literature review (part 1 of 3) (CSE Technical Report No. 731)*. Los Angeles, CA: University of California Los Angeles Center for the Study of Evaluation National Center for Research on Evaluation Standards and Student Testing(CRESST).