University of Groningen

The relation between teacher-set performance goals and students' mathematical achievement

Ritzema, Evelien S.; Deunk, Marjolein I.; Bosker, Roel J.; van Kuijk, Mechteld F.

Published in:
Studies in Educational Evaluation

DOI:
[10.1016/j.stueduc.2016.08.003](https://doi.org/10.1016/j.stueduc.2016.08.003)

IMPORTANT NOTE: You are advised to consult the publisher's version (publisher's PDF) if you wish to cite from it. Please check the document version below.

Document Version
Publisher's PDF, also known as Version of record

Publication date:
2016

[Link to publication in University of Groningen/UMCG research database](#)

Citation for published version (APA):

Ritzema, E. S., Deunk, M. I., Bosker, R. J., & van Kuijk, M. F. (2016). The relation between teacher-set performance goals and students' mathematical achievement. *Studies in Educational Evaluation*, 51, 17-28. <https://doi.org/10.1016/j.stueduc.2016.08.003>

Copyright

Other than for strictly personal use, it is not permitted to download or to forward/distribute the text or part of it without the consent of the author(s) and/or copyright holder(s), unless the work is under an open content license (like Creative Commons).

The publication may also be distributed here under the terms of Article 25fa of the Dutch Copyright Act, indicated by the "Taverne" license. More information can be found on the University of Groningen website: <https://www.rug.nl/library/open-access/self-archiving-pure/taverne-amendment>.

Take-down policy

If you believe that this document breaches copyright please contact us providing details, and we will remove access to the work immediately and investigate your claim.

Downloaded from the University of Groningen/UMCG research database (Pure): <http://www.rug.nl/research/portal>. For technical reasons the number of authors shown on this cover page is limited to 10 maximum.

The relation between teacher-set performance goals and students' mathematical achievement

Evelien S. Ritzema*, Marjolein I. Deunk, Roel J. Bosker, Mechteld F. van Kuijk

Groningen Institute for Educational Research, University of Groningen, Grote Rozenstraat 3, 9712 TG Groningen, The Netherlands

ARTICLE INFO

Article history:

Received 22 October 2015

Received in revised form 1 July 2016

Accepted 4 August 2016

Available online 28 August 2016

Keywords:

Teacher expectations

Goal setting

Academic achievement

ABSTRACT

The relation between teacher-set performance goals for 361 individual students and these students' mathematics achievement was investigated. High performance goals were found to strongly relate to student performance, with an effect size of $d = 0.80$. The performance goals were set by the teachers at the end of a step-by-step procedure, consisting of initial teacher expectations, the use of data, and team input. This procedure was expected to decrease negative expectancy bias. Higher teacher performance goals than teachers' initial expectations, so-called positive changes, were positively associated with the performance of initially low achievers. Initially high achievers, for whom the teachers made a positive change, performed worse than comparable students for whom initial expectation and final goal were the same.

© 2016 Elsevier Ltd. All rights reserved.

1. Introduction

Although the proficiency level of students in the Netherlands ranks comparatively high in international studies (Meelissen and Drent, 2008; Meelissen et al., 2012), there is concern about Dutch students' mathematical skills (Expert Group Continuous Learning Progression, 2008; Royal Netherlands Academy of Arts and Sciences, 2009). In particular, the low spread between the proficiency levels of the students, combined with the fact that the number of excellent students is decreasing (Meelissen et al., 2012), leads to doubts about whether teachers succeed in fitting their teaching to the varying needs of students and whether all students are challenged sufficiently.

In line with recent international educational policymaking, these concerns resulted in a focus on basic skills, on educational targets, and on data-driven decision making (DDDM). DDDM, or data use, can be used for school improvement purposes. Schildkamp and Kuiper (2010) defined data use as "systematically analyzing existing data sources within the school, applying outcomes of analyses to innovate teaching, curricula, and school performance, and, implementing (...) and evaluating these innovations" (p. 482). Through collecting, analyzing, and interpreting data – such as students' (standardized) test scores, results of informal assessments, daily work, information on classroom

observations, etc. – schools and teachers create knowledge on which decisions for adequate instructional adjustments can be based, thereby reflecting a formative use of data. Adjusting teaching to better fit student needs is assumed to lead to targeted and deliberate teaching, resulting in better student achievement. Goals play a key role in this way of working as they make clear what the desired level of student performance is (Lai and Schildkamp, 2013).

1.1. The use of data for instructional improvement

In order to support DDDM for instructional improvement purposes, a data-friendly school context can facilitate the effective use of data in class. Effective data use often presupposes changes in school culture and in the way schools and school teams handle information (Datnow, Park, & Wohlstetter, 2007; Huffman and Kalnin, 2003; Levin & Datnow, 2012). Teacher collaboration can support the individual and common knowledge and skills. Teachers can analyze and interpret the data together, diagnose areas in which students experience difficulties, give each other advice on instructional strategies, ask each other for help, and reflect on practice (Schildkamp and Kuiper, 2010; Wayman, Midgley, & Stringfield, 2006; Young, 2006). Therefore, professional development programs in DDDM should make use of collaboration (Wayman, Jimerson, & Cho, 2012). Wayman pointed in this respect to the advantages of small-group learning: the limited number of people could foster changes more easily and promotes active engagement of the participants. Moreover, by organizing

* Corresponding author.

E-mail address: e.s.ritzema@rug.nl (E.S. Ritzema).

professional development on data use in grade-level teams, program content and discussions can become more targeted.

DDDM is supposed to positively affect student achievement, provided that teachers make appropriate adjustments in teaching and instruction based on the data. Such adjustments should be mainly informed by one of the key features of data use: that schools and teachers make explicit what they strive for (Lai and Schildkamp, 2013; Wohlstetter, Datnow, & Park, 2008; Means, Padilla, & Gallagher, 2010). When appropriate measurable learning goals are set, goal attainment can be monitored in order to evaluate teaching. Information about student performance and daily work, among other things, provides teachers with feedback on their own teaching performance and such feedback can be used as guidelines for instructional improvement (Hattie and Timperley, 2007; Kluger & DeNisi, 1996). This ongoing process can be supported using reflective cycles, such as the formative models of Black and William (2009) and, in the Dutch context, the more test result-based approach of Ledoux, Blok, and Boogaard (2009).

In the Netherlands, teachers seem to make insufficient use of data for formative purposes: teachers do collect data (e.g., results on standardized tests and seatwork, and on students' classroom participation), but 50% of teachers refrain from taking instructional actions that result from their analyses (Educational Inspectorate, 2010). Evaluating performance seems to be problematic as well, as only 25% of schools were found to establish explicit goals, whereas such goals are necessary to evaluate student outcomes in terms of the desired levels of performance (Educational Inspectorate, 2010). A study on educators' ability to interpret results from a digital student monitoring system further showed that more than 70% of principals, senior support coordinators, and teachers did not reach the standard of accurate interpretation (that is, 85% correct on the items used in a questionnaire), with teachers performing significantly worse than the other two groups (van der Kleij and Eggen, 2013). Although ecological validity issues may play a role in interpreting these results (in daily practice, teachers might get support from the senior support coordinator in interpreting the results of the student monitoring system), this is deemed an undesirable situation. It seems to be unlikely that adequate instructional decisions can be made if they are based on incorrect interpretations.

1.2. Goal setting

The importance of goals is stressed in the literature on DDDM, because they reflect the desired achievement levels of a district, school, class, or individual student (Levin and Datnow, 2012; Carlson, Borman, & Robinson, 2011; Wohlstetter et al., 2008). In the current study, the sole focus was on performance goals. Performance goals are expected to have a positive effect on achievement, because people are highly motivated to reach them. The goal "give (s) an activity meaning or purpose" (Maehr & Zusho, 2009; p. 78) and, therefore, directs behavior. Locke and Latham (1990, 2002) argued that performance goals enhance effort, strengthen perseverance, and cause some sort of arousal. Especially in situations where people are committed to the goal, feel confident that they can attain it ("self-efficacy"), and receive feedback on their progress, goals are found to strongly influence performance (Locke and Latham, 2002).

Findings on the educational effectiveness of performance goals were reported by Fuchs, Fuchs, and Deno (1985). In their study on goal mastery and goal ambition, moderate to high ambitious goals were related to better student performance. Retrospective interviews showed that teachers felt more aware of the concrete goals. Assessment and evaluation information further seemed to facilitate instructional decision making in order to reach the goal. The findings of Fuchs and colleagues are in line with Locke and

Latham (2006), who found that people do not need to be consciously aware of the goal all the time, as they assume that "a goal, once accepted and understood, remains in the periphery of the unconsciousness as a reference point for guiding and giving meaning to subsequent mental and physical actions" (p. 267). Fuchs, Fuchs, and Hamlett (1989) further found that teachers who (regularly) adapted their performance goals according to students' actual performance were more ambitious than those who did not, which was related to higher student achievement. According to Fuchs et al. (1989), ambition thus seems to be inherent to the use of effective goals, although the authors also advocate that the goals should be attainable and not overambitious. Such realistic goal setting was promoted by providing the teachers in the experimental group information on the actual progress in student learning through performance information on several data points.

Performance goals that are set by teachers are probably to a large extent based on teachers' (tacit) beliefs and expectations in terms of their students' learning potential. By setting performance goals, teachers are requested to make these expectations explicit.

1.3. Teacher expectations

In a class, teachers are confronted with students who differ in terms of characteristics such as academic ability, concentration, motivation, effort, and classroom behavior. Teachers are aware of such differences and, as a result, develop implicit and explicit beliefs with respect to the class and its individual students. These naturally formed expectations can have an impact on aspects like teacher-student interaction, instructional decisions, opportunity-to-learn, materials used, and task difficulty (Rubie, 2003; Südkamp, Kaiser, & Möller, 2012; Jungbluth, 2012).

Teacher expectations can be held at different levels. At class level, teacher expectations can be high or low for the group as a whole. These general expectations at class level were found to be associated with differences in classroom climate, instructional decisions, and goal orientation (Rubie, 2003). At the student level, teachers have expectations of individual students in terms of their behavior, actual performance, and capabilities. Such individual expectancies may be influenced by the stereotypic beliefs a teacher has about certain groups. These expectations at group level have been a central topic in studies on inequalities in education (cf. for ethnic minorities: McKown and Weinstein, 2008 or Rubie-Davies, Hattie, & Hamilton, 2006; for social class: Ready and Wright, 2011; for sex: Jungbluth, 2012 or Tiedemann, 2002). Although teacher expectations can be biased by such group characteristics, Hattie (2009) and Good and Brophy (2003) both noted that the accuracy of the expectations improves once teachers are given information about prior achievements or students' participation in academic activities. A more refined picture of the individual student thus leads to more accurate expectations and such accurate beliefs about a student's capabilities enable teachers to make adequate instructional decisions (Martínez, Stecher, & Borko, 2009). A salient reason for avoiding teacher expectation bias is that a substantial part of such bias remains stable in time (de Boer et al., 2010; Boer, Bosker, & van der Werf, 2010). Schools and teachers should thus become aware of their expectations and should aim to avoid being too hesitant or careful about what they expect from their students, as this would reflect negative expectation bias. A realistic and accurate view of what students are capable of is to be promoted, or perhaps even a somewhat (too) optimistic and ambitious expectation, reflecting positive expectation bias. The promotion of ambitious expectations is important, as positive expectancy bias has been found to positively affect student performance (de Boer et al., 2010). Hence, de Boer and colleagues (2010) advocated that "... schools and teachers should be more

aware of the positive effect of positive expectation bias and that they might maximize their expectancies, within certain limits, for each student" (p. 178).

There is empirical evidence that teacher expectations and student achievement are related. Findings from educational effectiveness studies provide evidence that high teacher expectations are associated with better student performance (Scheerens & Bosker, 1997; Scheerens & Bosker, 1997; Sammons, Hillman, & Mortimore, 1995). These studies aimed at determining the key characteristics for effective teaching, thereby being comprehensive in nature. When specifically looking at research on teacher expectations, Jussim and Harber (2005) concluded that the effects of teacher expectancies, reflected through self-fulfilling prophecies, tend to be small. Self-fulfilling prophecies can be understood as relatively high or low teacher expectations when compared to the actual capacities of the students, yielding better/worse student performance. Such self-fulfilling prophecy effects were found for approximately 5% to 10% of the students, whereas for most students, teacher expectations tended to be accurate. Using "prediction without causation" to define accuracy (Jussim & Harber, 2005; p. 141), Jussim and Harber argued that for most students teacher expectations predict but do not cause or influence their performance.

Findings from more detailed studies, focusing on how the teacher expectations were measured, refined Jussim and Harber's (2005) conclusions somewhat. A factor that seemed to influence accuracy was the use of informed or uninformed judgment. Südkamp et al. (2012) found that judgments on student achievement were more accurate when information from standardized tests was used, as opposed to more informal judgments, in which teachers were unaware of the standard of comparison. Both the studies by Begenly, Eckert, Montarello, and Storie (2008) and Feinberg and Shapiro (2009) showed that the accuracy of teacher judgments depended on whether teachers had to compare and list their expectations of a student relative to other students in class (the relative performance level) or whether they had to predict how well students would score on a test (the absolute performance level). Teachers have been found to do fairly well on relative judgments (Hoge and Coladarci, 1989; Südkamp et al., 2012), whereas predicting student performance in absolute terms is more difficult for them. In predicting absolute achievement, differential effects were found: teachers tend to be more accurate for the good or average performers than they are for low-achieving students (Begenly et al., 2008; Feinberg & Shapiro, 2009).

The transformation of teacher expectations into explicit teacher-set performance goals might help to avoid negative expectancy bias (Bates and Nettelbeck, 2001; Feinberg & Shapiro, 2009). By deliberately thinking about students' capabilities, while taking into account several sources of data that could inform this process – for instance, performance data or knowledge shared within the school team –, teachers might modify their beliefs about some or all students. In setting goals, teachers are further assumed to become aware of their own potential role in developing students' abilities. After all, the teachers work with the students and can make efforts to optimize student learning, for example, by providing more challenging questions, tasks, and interactions, that is, by offering them more opportunity-to-learn (McDonald et al., 2016; Rubie-Davies, 2007).

1.4. Research questions

There is empirical evidence for the assumption that teacher expectations and performance goals are associated with better student outcomes (Fuchs et al., 1985; Scheerens & Bosker, 1997). Teacher expectations seem to become more accurate and well-

considered if more information sources are taken into account, such as, performance data. As a result, negative expectation bias may be minimized, and positive expectation bias may be deliberately introduced. The introduction of positive expectation bias is in line with de Boer et al. (2010) de Boer and colleagues (2010), who promoted maximized teacher expectations and positive expectation bias, which would ultimately foster student opportunities. In the current study, we aimed to investigate whether we could relate teacher-set performance goals to student achievement, more specifically, in situations where expectancy bias was assumed to be minimized due to a step-by-step procedure consisting of both performance information and team input. This led to the following questions:

1. To what extent are well-considered performance goals related to students' mathematics achievement, while controlling for pretest information and general academic ability?
2. To what extent do teachers' well-considered performance goals differ from teachers' initial expectations?
3. If they differ, to what extent are 'positive changes' associated with students' mathematics achievement, while controlling for initial differences? And to what extent can differential effects of the 'positive changes' be determined for the initially low and initially high mathematics performers?

Our first hypothesis was that there would be differences between the performance goals and the preliminary expectations, indicating that teachers used the information given in a step-by-step procedure at the end of which teachers had to set the final goals (Good & Brophy, 2003; Hattie, 2009). This procedure was assumed to stimulate teachers' use of several data sources, thereby refining possibly incorrect or biased preliminary teacher expectations. Furthermore, we expected that these performance goals related to student achievement, and that high teacher goals were associated with better mathematics achievement of the students (Hypothesis 2). This hypothesis resulted partly from our assumption that teachers would perform better as a result of their enhanced motivation to reach the goals they had set. In their attempts to reach their goals, teachers were supposed to make more deliberate instructional decisions in order to better fit their teaching to students' needs, for instance, through improved questioning techniques. Such a better adjustment of instruction to student needs is expected to foster students' learning and achievement. Another reason for expecting improved student outcomes was the assumption that setting and monitoring goals improves the quality of the reflective cycles used to work in a data-driven way. Given the deliberate decisions and improved teacher reflection it was expected that positive changes, meaning that the final goal was higher than the initial expectation and thus reflecting minimized negative or even positive expectancy bias, would be positively related to student achievement (Hypothesis 3).

2. Method

2.1. Design

To determine the general relation between (high) teacher-set goals and students' mathematics achievement, a pre-posttest design was used, in which the outcome measure was a standardized mathematics assessment. The relation between a positive change – teachers setting a higher final goal than their initial expectation – and students' mathematics achievement was investigated using a pre-posttest control group design. The control group was constructed through propensity score matching.

2.2. Participants

Student assessment data and teacher-set performance goals were collected for 361 students. Of these 361 students, 176 students were students in grade 2 and 185 were students in grade 3. There were 196 boys and 165 girls. The 361 students were taught in 27 classes. More than half of the 27 classes were multi-grade classes, consisting of two or more grades per class. The 27 classes were taught by 35 teachers, some holding part-time positions. Teachers teaching the same class were asked to set the performance goals collaboratively. Teachers had, on average, 16.3 years of experience ($SD = 10.7$), and the majority of the teachers was female (94%). The teachers who participated in the current study were all involved in a professional development program (PDP) on data use.

2.3. Procedure

Performance goals and standards were elements of a PDP on DDDM, consisting of three core, interrelated components: (a) performance goals and standards, (b) data-analysis, and (c) instruction. At the start of the PDP, the teachers were asked to reflect on what students should be able to know and do in grades 2 and 3 and, more specifically, on how the students in a specific grade differ in terms of their mathematics proficiency. For this purpose, the teachers were asked to set performance standards. Performance standards are supposed to provide information on students' levels of mastery of content standards (Cizek & Bunch, 2007; Hambleton & Pitoniak, 2006; Linn & Baker, 1995). Such content standards determine what knowledge and skills should be taught to the students at different grade levels, reflected in the curriculum. As the performance standards indicate to what levels the content standards are met (for instance, whether students master the content standards in an acceptable or outstanding way) they are closely linked to assessments. The performance standards are operationalized in terms of cut scores on the proficiency scale of a specific assessment.

During a mathematics standard-setting procedure (SSP), teachers were asked to reflect on the desired knowledge and skills that students of grades 2 and 3 were supposed to master at the end of the school year. The teachers were requested to set performance scores for students of four different performance levels (*minimum*, *basic*, *proficient*, and *advanced*). In order to accomplish this task, the teachers had to consider an ordered-item booklet consisting of 30 standardized assessment items (a selection of items of the mathematics standardized assessments used in grades 1–4) and decide for each of the four performance levels which assessment items the 'just competent student' should be able to answer correctly (Mitzel, Lewis, Patz, & Green, 2001). For an elaboration on standards and the standard-setting procedure used in the PDP, see Deunk, van Kuijk, and Bosker (2014). The standard-setting procedure resulted in four standards per grade. By taking the intervals between these standards, five performance categories were established (below minimum, minimum, basic, proficient, and advanced). In total, 10 performance categories were specified, five for grade 2 and five for grade 3. These five performance categories per grade were used to determine the performance goals, as the teachers were asked to assign a goal, reflecting one of the five grade-specific performance categories, for each of their students.

To ensure that the performance goals were well-considered, we designed a step-by-step procedure consisting of the following four steps: (1) determining an initial expectation, (2) the use of performance data, (3) collaboration, and (4) assigning a final goal. In step 1, teachers had to make their expectations explicit for each student in terms of one of the five performance categories (initial

expectation). In step 2, information on performance data was used. Teachers were asked to compare performance information from the student monitoring system with their initial expectations. For this task, teachers were asked to use the same five performance categories as used in step 1 to record their findings. Comparing initial expectations with the actual performance data was expected to stimulate the teachers' reflection on their initial expectations, particularly as they were requested to answer questions regarding observed differences between step 1 and step 2. Step 3 consisted of team collaboration. Steps 1 and 2 together formed a homework assignment that each teacher had to complete prior to the meeting in which the final goals were to be set. During this meeting, the homework assignments were discussed in the school team. These discussions served to stimulate teachers' reflection, as colleagues within the school – teachers, the senior support coordinator, and the school principal – questioned some of the teachers' expectations and added relevant information on, for instance, specific instructional or socio-emotional approaches in relation to specific students. The fourth and final step was assigning a final goal. After the group discussion, teachers had to take into account all the information obtained, reconsider their initial expectations, and set a performance goal for each individual student. As we assumed that making expectations explicit would involve some level of ambition and as, empirically, effective performance goals were found to reflect at least some ambition (Fuchs et al., 1989), teachers were asked to set a challenging but realistic goal for each student. The performance goals for the standardized assessment in May/June were set in fall, leaving the participants with approximately six to seven months to reach them.

2.4. Variables and instruments

2.4.1. Mathematics assessment

The Cito standardized assessments for mathematics are part of the Cito LOVS assessment system that is used in 85% of Dutch schools, including the schools that participated in the PDP. The tests are used throughout primary school (grades 1–6), and students are assessed twice a year: halfway through the school year (mid-term assessment, January) and at the end of the school year (end-of-the-year assessment, May/June). Test scores for all assessments (from grade 1 to 6) are associated with a single proficiency scale, ranging from 0 to 169. The assessments have a good overall reliability in all grades (Cronbach's alpha is at least 0.91, Janssen, Verhelst, Engelen, & Scheltens, 2010). Test scores before the start of the PDP were used as pretest (May/June 2011) and test scores at the end of the PDP as posttest (May/June 2012).

2.4.2. Reading comprehension assessment

The Cito standardized assessments for reading comprehension (RC) are used in grades 1–6 in primary school and are also part of the Cito LOVS assessment system (for more information on the Cito LOVS assessment system the reader is referred to van der Kleij and Eggen, 2013; van der Kleij, Eggen, & Engelen, 2014). The (arbitrary) proficiency scale ranges from –87 up to +147. The tests have a good overall reliability (Cronbach's alpha is at least 0.89, Feenstra, Kleintjes, Kamphuis, & Krom, 2010). Students' scores on the RC assessment were taken into account, as student achievement on mathematics is likely to be related to more general capabilities, and teachers probably considered such general capacities during the goal setting process. In the analyses, pretest scores on the RC assessment were, therefore, used as a proxy for general academic capability.

2.4.3. Heterogeneity of the class

The heterogeneity of the class, meaning the spread in student achievement levels within the class, was taken into account

regarding mathematics and reading comprehension. For class heterogeneity on mathematics, the standard deviation of the class's mathematics pretest scores was used; for the class heterogeneity on reading, the standard deviation of the class's RC pretest scores was used.

2.4.4. Initial expectations

At the start of the step-by-step goal setting procedure, teachers were asked to write down their expectations for each student. These initial expectations were set in terms of five performance categories. These categories, five intervals on the proficiency scale of the posttest, reflected the intervals between the four performance standards. The five performance categories – below minimum, minimum, basic, proficient, and advanced – were coded 1–5.

2.4.5. Performance goals

Teachers were asked to set a performance goal for each individual student in terms of the five performance categories described above. These performance goals reflect the final goals, step 4 in the step-by-step procedure.

2.4.6. Positive change

At the start of the multistep procedure, teachers had to write down their initial expectation. At the end of this procedure the teachers were asked to set a final goal. If the final goal was higher than the initial expectation, this was called a positive change. Positive change was a dummy-coded variable ("no change" was the reference group).

2.4.7. Sex

Ganley et al. (2013) refer to various studies in which a (small) sex difference in mathematics test performance was found, with boys outperforming girls on tests. Sex, a dummy-coded variable with boys as reference group, was, therefore, used as a covariate.

2.4.8. Grade

The PDP focused on teachers in grades 2 and 3. To account for performance differences between these two grades, grade was used as a covariate as well. Grade was a dummy-coded variable, where grade 2 was the reference group.

2.5. Analyses

For answering the first question, concerning the relation between teachers' (high) performance goals and students' mathematics achievement, a multilevel multiple regression analysis was carried out, using MLwiN software (Rasbash, Browne, Healy, Cameron, & Charlton, 2012). Two levels were distinguished: Students were considered to be nested in classes. Three models were estimated: the Start model, containing only the math pretest, the Covariate model, controlling for mathematics pretest, RC pretest, sex, and grade, and the Main Effect model, in which the final goal was added. Beforehand, checks on the general levels of ambition of the teachers were carried out, investigating the between-teacher differences in terms of the overall height of the teachers' goals for their students. No differences were found in this respect.

The second question, investigating whether teachers changed their initial expectation at the end of a step-by-step goal-setting procedure, was answered using descriptive analyses.

The third question, regarding (a) the general relation between student achievement and a positive change and (b) the differential effects of the positive change for initially low or high performers, was also investigated using two-level analyses. Again, three models were estimated. First, a Covariate model was estimated,

controlling for mathematics pretest (grand-mean centered for ease of interpretation). Next, the positive change was added to estimate a Main Effect model. Thereafter, adding the interaction *Goal*Math pretest* yielded the Interaction model.

2.6. Construction of a control group

To answer the third question, two comparable groups were constructed, consisting of (a) students for whom teachers changed their initial expectations to a more positive final goal—the "positive change" group—and (b) students for whom the final teacher goal was the same as teachers' initial expectation—the control group. This was done using Propensity Score Matching (PSM), a matching method by which the selection process is modeled (Rosenbaum & Rubin, 1985). As such, causal inference is allowed for. By comparing the *positive change students* with students for whom initial expectation and final goal were the same but who were similar to the positive change students in terms of important observed characteristics, these two groups were considered to be comparable except for having a raised final teacher goal or not. More formally, by using PSM the distribution of the observed covariates in the positive-change group and the control group was balanced in such a way that students with the same propensity scores had the same distributions over all covariates, without the need for exact matches on the individual variables used (Rubin, 2007; Stuart, 2010). The propensity score (PS) then summarized all the information that predicted the positive change. In the current study, this PS could be defined as the probability that a student would receive a higher final goal than the initial expectation, given the set of observed variables.

In conducting the PSM, we selected the 0.2 caliper matching without replacement algorithm. This matching algorithm only matched positive-change students to control students within a certain range (0.2 standard deviations), finding the best match in terms of the PS. In this way, inadequate matches were avoided and, thereby, selection bias was decreased (Caliendo & Kopeinig, 2008). A further advantage of building in such a tolerance level is that it supports meeting the Common Support assumption, meaning that there was sufficient overlap in the characteristics of positive-change students and control students. Using the caliper matching algorithm entailed that students were left out for whom no match could be found within the tolerance level.

2.7. Variables used for propensity score matching

A crucial assumption of PSM is that, conditional on the observed variables used, treatment assignment is independent of the potential outcomes, meaning that there should be no hidden bias or confounding (Stuart, 2010). The selected observed pre-treatment variables should further be associated with both treatment assignment and outcome (Brookhart, Schneeweiss, Rothman, Glynn, Avorn, & Stürmer, 2006). Steiner, Cook, Shadish, and Clark (2010) stressed in this respect the importance of including matching variables such as baseline test scores, given their strong relation to the outcomes. Taking such baseline measures into account seems to be particularly salient for small samples (Brookhart et al., 2006). Because of the relatively small sample in our study, we matched on nine variables, most of them related to the outcome. Five variables at class level were used: (1) grade, (2) the average mathematics pretest score of the class, (3) class heterogeneity on the mathematics pretest, (4) the average RC pretest score of the class, and (5) class heterogeneity on the RC pretest. Four student level variables were taken into account: (6) sex, (7) teacher's initial expectation, (8) the mathematics pretest score, and (9) the RC pretest score.

Table 1

Summary of Students' Test Results on the Mathematics Assessment, Pretest, and Posttest for Second and Third Grade.

Grade	<i>n</i>	Pretest Mean (<i>SD</i>)	Posttest Mean (<i>SD</i>)
2	176	48.8 (14.9)	67.4 (15.8)
3	185	63.4 (15.0)	80.0 (14.5)
Total	361	56.3 (16.6)	73.9 (16.4)

Note. For all assessments in the Cito LOVS assessment system (grade 1–6) one proficiency scale is used.

Table 2

Overview of the Differences between Post-Test Scores and Final Goals.

Difference between post test score and final goal	<i>n</i>	%
Posttest two categories lower than final goal	4	1.1
Posttest one category lower than final goal	71	19.7
Category posttest and final goal is the same	210	58.2
Posttest one category higher than final goal	66	18.3
Posttest two categories higher than final goal	9	2.5
Posttest three categories higher than final goal	1	0.3

3. Results

The mean pre- and posttest mathematics scores of the students in grades 2 and 3 are depicted in Table 1. From the data in Table 1 it can be derived that the average increase in the scores of the students in grade 2 was 18.6, whereas it was 16.6 for students of grade 3.

Teachers were asked to set a challenging but realistic final goal for each of their students in terms of five performance categories. As each performance category reflected ranges of expected posttest scores, students' actual posttest scores could be translated to these five categories as well. It was found that almost 60% of the students had a test score in the same category as their final goal, approximately 20% attained a score that was higher than their final goal, and approximately 20% of the students attained a score that was lower than their final goal (see Table 2).

In order to investigate the general relation between the final goals and student achievement, a multilevel analysis was used. In Table 3, three models are presented, containing standardized regression coefficients.

In the Start model, the mathematics pretest scores were found to significantly predict the mathematics posttest scores (see Table 3). In the Covariate model the other covariates were added as well. From the Covariate model it can be derived that, next to the mathematics pretest score, sex and the RC pretest score significantly predicted the mathematics posttest scores, but grade did not. Deviance testing indicated that the Covariate model fitted the data better than the Empty model ($\chi^2 = 29.59$, $df = 3$, $p < 0.001$). Adding "final goal" (the Main Effect model) further significantly improved model fit ($\chi^2 = 57.67$, $df = 1$, $p < 0.001$). These findings could be interpreted as follows: keeping all covariates equal, a rise in 'goal' by one category which, coincidentally, equaled the standard deviation of the final goal, is associated with an increase of 0.37 *SD* on the posttest score. Thus, when different final goals were assigned to two apparently similar students (based on the covariates used), the higher goal was associated with a substantially and significantly higher achievement on the mathematics posttest, with a corresponding effect size of Cohen's $d = 0.80$.¹ Adding the goals accounted for 15% of the variance in mathematics posttest scores.

The final goals were set by the teachers at the end of the step-by-step procedure that aimed at stimulating teachers' reflection on their own initial expectations. The final performance goals were set with respect to five performance categories. The range of ability scores on the posttest per desired performance category (that is, per final goal) is depicted per grade in Table 4, as well as the number of students per final goal. The mean of the teacher-set final goals was 3.74 ($SD = 1.06$); on average, teachers expected the student achievement level to lie between the basic and the proficient level.

As a result of the step-by-step procedure, the performance categories for a specific student's initial expectation and for this same student's final goal could differ. Table 5 shows how the final goals differed from the initial expectations. Initial expectations were available for 276 students. For most of these students ($n = 198$), the final goal was the same as the initial expectation, but for 78 students the initial expectation was adapted. Teachers became more ambitious (a final goal that was one or two categories higher) for 70 students; for only eight students the teacher expectations were tempered (the goal was one or two categories lower).

The multilevel results depicted in Table 3 showed that higher goals were associated with higher student achievement, while controlling for initial differences between students. Nevertheless, from this result it cannot be concluded whether the higher goals reflected a positive expectancy bias or whether they were realistic. The positive changes that resulted from the step-by-step goal setting procedure are assumed to provide more information on the ambition of the goals, as they indicate teachers' increased ambition for certain students. To investigate whether such a shift towards a more positive final goal related to student achievement, the students for whom the performance goals were higher than the initial expectations were compared to similar students for whom the initial expectations and the performance goals were the same.

Using PSM, students for whom the final goal was higher than the initial expectation ($n = 70$) were matched to comparable students. The latter students came from a pool of 198 possible controls, for whom the final goal was the same as the initial expectation.² Students were matched based on both class- and student-level characteristics. Matching quality was assessed by checking similarity on the covariate distribution in the two matched groups, using independent samples *t*-tests. The procedure should yield no remaining significant differences between the two groups of students after the matching. In Table 6 the mean scores of both the positive-change group and the control group are shown, both before and after matching. Before matching, the groups differed in terms of the variables sex and initial expectation, as indicated by the asterisk. The positive-change group consisted of more boys than girls, whereas the number of boys and girls was comparable in the control group. Further, the average initial expectation of the positive-change group was significantly lower than that of the control group. After the matching, the groups no longer differed significantly on all nine variables. The matching procedure led to a total of 114 students, 57 in the positive-change group and 57 in the control group.

The matching quality was also investigated by checking whether the Common Support assumption was met, meaning that there should be sufficient overlap between members of both matched groups in terms of their propensity score. From Fig. 1 it can be derived that the propensity scores of the two groups sufficiently overlapped. Moreover, the region of common support

² Students for whom teachers set a final goal that was lower than the initial expectation were excluded from the matching procedure and, consequently, from the subsequent multilevel analyses.

¹ In calculating the effect size the following formula was used: $d = 2r/\sqrt{1 - r^2}$.

Table 3

Results from Multilevel Regression Models of Performance Goals.

	Start model		Covariate model		Main Effect model	
	β	SE	β	SE	β	SE
Response variable: math posttest						
<i>Fixed part</i>						
Intercept	32.27	2.06	39.63	2.39	29.94	2.53
Math pretest	0.75*	0.03	0.62*	0.04	0.35*	0.05
Sex			−0.15*	0.03	−0.12*	0.03
Grade			0.00	0.04	0.21*	0.05
Reading pretest			0.19*	0.04	0.11*	0.04
Final goal					0.37*	0.05
<i>Random part</i>						
Class level	4.17	3.35	2.76	2.75	2.74	2.47
Student level	107.57	8.27	99.82	7.68	84.82	6.52
−2loglikelihood		2724.09		2694.50		2636.83
No of students		361		361		361

Italics: standardized coefficients.

* $p < 0.05$.**Table 4**

Frequencies of the Final Performance Goals per Grade.

Final goal	Grade 2		Grade 3	
	Range of ability scores	No. of students	Range of ability scores	No. of students
Below Minimum	≤ 36	3	≤ 56	7
Minimum	37–47	18	57–66	21
Basic	48–60	34	67–76	51
Proficient	61–70	69	77–91	60
Advanced	≥ 71	52	≥ 92	46
Total		176		185

was relatively large, with propensity scores ranging from close to zero to almost 0.9.

The PSM led to two comparable groups of 114 students in total, being taught in 23 classes. The mean pretest score of these 114 students was 56.7 ($SD = 15.5$). For the matched positive-change students ($n = 57$), pretest scores ranged from 24 to 86; for the matched control students ($n = 57$), pretest scores ranged from 14 to 109. In Table 7, the mean scores for each initial-expectation category are presented. Note that the nine students who received an initial expectation in the advanced category all belonged to the control group; positive change students could, by definition, not start in the highest category.

For answering the third research question, concerning the relation between positive changes and student achievement, multilevel analyses were used, regressing the mathematics achievement of the students (the posttest scores) on the positive change. Three models were estimated: a Covariate model, a Main Effect model, and an Interaction model. From the results presented in Table 8 it can be derived that the pretest – individual mathematics scores – significantly predicted the posttest. When

adding positive change (the Main Effect model), no significant effect of positive change was found: deviance testing indicated that the main model did not fit the data better than the covariate model ($\chi^2 = 0.03$, $df = 2$, $p = 0.99$). Next, a differential effect of the positive change for initially low or initially high mathematics performance was investigated by adding an interaction term to the Main Effect model. This Interaction model showed a significant negative interaction effect. The deviance decreased significantly, indicating that the Interaction model fitted the data better than the Main Effect model ($\chi^2 = 6.91$, $df = 2$, $p = 0.03$). Positive changes were thus differently related to initially low- or high-performance.

In Fig. 2 it is shown how this negative interaction effect should be interpreted. On the X-axis, the ability scores on the pretest are reported and on the Y-axis the ability scores on the posttest. Fig. 2 contains two lines. The dark line represents the estimated posttest scores for students for whom the teachers had set a positive change and the dashed line represents the estimated posttest scores for the students for whom there was no change. As can be derived from Fig. 2, the crossing lines indicate that the positive changes influenced the posttest scores differently for initially low and high

Table 5

Overview of the Initial Expectations and the Teacher-Set Final Goals.

Initial expectation	Final goal					Total
	Below Minimum	Minimum	Basic	Proficient	Advanced	
Below Minimum	8	3	4	0	0	15
Minimum	1	26	14	1	0	42
Basic	0	3	41	25	1	70
Proficient	0	1	2	77	22	102
Advanced	0	0	0	1	46	47
Total	9	33	61	104	69	276

Table 6

Student and Class Characteristics in Positive Change Group and Control Group, before and after Matching.

	Before matching				After matching			
	Positive change (n = 70)		Control group (n = 198)		Positive change (n = 57)		Control group (n = 57)	
	M	SD	M	SD	M	SD	M	SD
<i>Student level</i>								
Initial expectation	2.90*	0.97	3.64	1.10	3.09	0.85	3.14	1.19
Pretest math	55.50	14.28	56.69	16.70	56.47	15.32	56.95	15.76
Pretest RC	10.61	17.89	13.41	17.97	11.72	18.45	11.88	16.47
Sex	0.34*	0.48	0.49	0.50	0.35	0.48	0.39	0.49
<i>Class level</i>								
Grade	0.63	0.49	0.53	0.50	0.65	0.48	0.61	0.49
Heterogeneity math	13.53	3.10	14.18	3.03	13.38	3.19	13.36	3.00
Class's mean math score	56.48	9.36	56.04	8.35	57.61	9.42	58.29	7.36
Heterogeneity RC	12.87	11.26	13.70	4.33	13.51	4.23	13.62	3.88
Class's mean RC score	13.38	3.85	13.69	4.33	13.07	12.04	12.69	7.99
Propensity Score					0.36	0.19	0.35	0.19

* $p < 0.05$. **Fig. 1.** Distribution of Propensity Scores.

performing students. The initially low performing positive change students were found to perform better than the initially low performing students for whom no change was made by the teacher, whereas the contrary was found for the initially high achieving students. Initially high achieving students for whom initial teacher expectations and final goals were the same were found to perform better on the mathematics posttest than the positive change students. The crossing point of the lines is at 57.8, which is relatively close to the average mean score on the pretest ($M = 56.7$).

3.1. Checks for robustness of the findings

In conducting the matching, the algorithm 0.2 caliper matching without replacement was used. However, as Caliendo and Kopeining (2008) note, ‘asymptotically all PSM estimators should

yield the same results’ (p. 44). In order to check for the robustness of the PSM results, seven additional algorithms were used.³ Five of these yielded a significant negative interaction effect: Initially low performers profited from positive changes, whereas initially high performers did not. This means that in total six out of eight PSM algorithms resulted in the same impact of positive changes for low and high achievers. The two deviating results came from the 0.2 caliper matching with replacement, using both class- and

³ The following seven algorithms were used: (a) 0.2 caliper with replacement using class and student variables, (b) 0.2 caliper with replacement using student variables, (c) 0.2 caliper without replacement using student variables, (d) 0.1 caliper without replacement using class and student variables, (e) nearest neighbor without replacement using class and student variables, (f) nearest neighbor with replacement using student variables, and (g) nearest neighbor without replacement using student variables.

Table 7
Summary of the Mean Pretest Scores per Initial Expectation Category.

Category of initial expectation	<i>n</i>	<i>M</i>	<i>SD</i>
Below minimum	6	33.0	11.5
Minimum	27	47.5	8.6
Basic	38	53.8	11.5
Proficient	34	65.4	11.3
Advanced	9	79.6	16.2

Fig. 2. Posttest estimates for the initially low and high performers.

student-level variables and 0.1 caliper matching without replacement, using both class- and student-level variables. The first matching algorithm led to a significant positive main effect of the positive change and no significant interaction effect (positive-change group: $n=63$; control group: $n=46$). The 0.1 caliper matching algorithm, with a smaller tolerance level, did not support a significant interaction effect. However, the interaction effect was only just insignificant and analyses were conducted based on 53 students in both the positive-change and the control group. The lack of significance might, therefore, partly be caused by somewhat lower statistical power.

4. Conclusion and discussion

In the current study it was investigated whether relations could be identified between well-considered teacher-set performance goals, positive changes, and students' mathematics achievement. Our second hypothesis (H2) was that higher, well-considered goals were associated with higher student performance. The results of the multilevel analysis confirmed H2, as setting higher performance goals indeed related to better mathematics achievement of

the students, with an effect size of $d=0.80$. This could be considered a strong relation, in terms of Cohen's classification.

At the end of a step-by-step procedure, the performance goals were set in terms of five performance categories. The assumption was that effective goals are realistic yet challenging (Fuchs et al., 1989; Locke & Latham, 2002). Results on the attainment of the goals showed that a large number of students (approximately 80%) scored at or above their desired performance level. This high percentage of students attaining the goals seems to imply that the goals were realistic (that is, the fast majority of the students reached the aim set by the teacher), but at the same time this high percentage makes it questionable whether the teachers set the goals for their students in an ambitious way. Descriptive analyses of the goal setting procedure indicated that teachers retained their initial expectations for most students ($n=198$). Yet for more than a quarter of the students (28%), teachers changed their initial expectations, indicating that the use of two data sources, performance data and team input, triggered further reflection on students' capacities. These findings suggest that teachers used the information provided by the step-by-step procedure, which is in line with our first hypothesis. Most of these changes led to a higher final goal than the initial expectation (for 70 of the 276 students for whom initial expectations were available). It was expected that such positive changes indicated either minimized negative expectation bias, meaning that inaccurate low teacher expectancies were adjusted, or positive expectation bias, suggesting that teacher expectancies became a little bit too positive.

The third research question concerned the relation between positive changes and students' mathematics achievement, and the differential effects of such positive changes for initially low or high mathematics performers. For this purpose, the students for whom there was a positive change were matched, using propensity score matching, to comparable students for whom the initial expectation and the final goal were the same. This led to two similar groups, containing 57 students in each condition. Impact analyses were conducted using multilevel analyses. Although no main effect of the positive change was found, a negative interaction effect was identified. This negative interaction effect indicated that the positive change influenced the posttest scores of initially low performers in a positive way, but negatively affected those of initially high performers. The hypothesis that positive changes would lead to better student achievement was thus only partially confirmed.

4.1. Interpretation of the results

In response to question 1, a strong relation between (high) performance goals and students' mathematics achievement was found. However, we should refine this finding somewhat. Although

Table 8
Results from Multilevel Regression Models of Positive Change.

	Covariate model		Main model		Interaction model	
	β	SE	β	SE	β	SE
Response variable: math posttest						
<i>Fixed Part</i>						
Intercept	75.11	1.25	74.93	1.60	74.82	1.57
Pretest (gmc)	0.86*	0.07	0.86*	0.07	1.02*	0.09
Positive change			0.37	2.01	0.40	1.95
Positive change $\times$ pretest (gmc)					-0.33*	0.13
<i>Random Part</i>						
Class level	8.85	9.13	8.94	9.16	9.73	9.16
Student level	107.92	15.45	107.83	15.44	100.67	14.44
-2 $\times$ loglikelihood:	864.36		864.33		857.42	

gmc: grand-mean centered.

we aimed to take into account important information on initial differences between students (mathematics performance levels and general academic capabilities), teachers probably also took other student characteristics, such as their daily work, classroom behavior, motivation, effort, concentration, and self-esteem, into account while setting the goals. Given the assumption that the teacher expectations incorporated more features than only academic achievement, it is difficult to draw conclusions from the general goal-achievement relation on the ambitiousness or accuracy of such high goals. Slavin (1990) argued in this respect that additional teacher knowledge might lead to the accurate expectation that the mathematics potential of two students may well differ despite their identical scores on standardized cognitive pretests. Goals that were considered high in our first multilevel analysis might not be so high in reality and, therefore, might not reflect high expectations. Focusing on ‘positive changes’ was assumed to provide more information on the issue of teacher ambition. As teachers were asked to state their initial expectations at the beginning of the goal setting procedure, it was probable that they took all kinds of student characteristics into account at that point. During the step-by-step procedure, this initial expectation was refined, leading to higher final goals for a quarter of the students. Investigating the relation between such positive changes and student performance was assumed to provide more information on whether a higher level of ambition, or perhaps a lower level of negative expectation bias, would benefit students’ achievement.

The Interaction model showed that initially weak performers profited from the positive change in that they achieved better, whereas the initially stronger students were found to score lower. Two underlying mechanisms might have played a role here. It is possible that the step-by-step procedure led to a decrease in negative expectation bias for the low achieving students. As a result, the final goals for these students could be considered more accurate and teachers’ instruction might have better suited students’ actual performance levels. This conclusion would be in line with the findings of Begeny et al. (2008) and Feinberg and Shapiro (2009) who observed that teachers were less accurate in predicting absolute achievement for weak performers than for the average and high performers. For the (average and) advanced students, the teachers in our study may have deliberately assigned goals that were too ambitious, indicating positive expectation bias, while overlooking the need for extra help that would enable these students to actually perform at the higher level. The instruction might, therefore, have fitted the needs of these students to a lesser degree. An observational study of the same teachers’ differentiation behavior might improve understanding of the counterintuitive finding that positive changes in expectations negatively affected the outcomes of high performers (Ritzema, Deunk, & Bosker, 2016). Observations that were conducted near the end of the school year showed that teachers mostly addressed struggling students, but did not address advanced students in providing additional (challenging) instruction. Setting higher goals than their initial expectations for their average or advanced students might thus have caused the teachers to give these students more challenging tasks without providing the additional support that might have been needed for the students to master these tasks. Consequently, although the teachers became more ambitious in their goals for the average and advanced students, the ineffectiveness of these positive changes for these students could have depended on the way the teachers acted on or refrained from acting on these ambitious expectations.

Naturally, positive shifts in final goals do not necessarily reflect a shift from negative expectation bias to neutral, accurate expectations for low performers. Another possible explanation is that the teachers, having their more ambitious goal in mind, provided more opportunities to learn to these low achievers, for

instance, by giving them additional instruction, by informally assessing their knowledge more often, by giving them more feedback, by encouraging them more, or by letting them answer questions more often (Good & Brophy, 2003; Rosenthal, 1994). The (subtle) ways in which these positive expectations tend to be communicated are expected to lead to better student outcomes. The fact that particularly low achievers profited from such modified teaching is in line with Madon, Jussim, and Eccles (1997), who found that both positive and negative expectations were more strongly related to the performance of low-achieving students than to that of high achievers. Low achievers are, therefore, assumed to be more susceptible to high or low teacher expectations.

Whatever underlying mechanisms (reduced negative expectation bias or positive expectation bias, or both) could explain the findings of this study, the positive changes led to convergent differentiation, meaning that differences between student achievements became smaller. The increased achievement of the struggling students might have been related to the efforts of the teachers, who might have raised the bar for and adjusted their instruction to better suit their teaching to these students’ needs. However, following this line of reasoning, everything seems to have its price. Increasing attention to the low-achieving students possibly prevented teachers from making sure that the high achievers for whom they had positive expectations were sufficiently supported, possibly due to time constraints.

4.2. Limitations

Three limitations to our study should be addressed. A first limitation was that no information was obtained on whether and how setting the goals had been translated by the teachers into modifications of classroom instruction. We thus do not know how working with goals was implemented by the teachers, although such information would have helped to understand our findings. As the subtle ways of communicating teacher expectations might be difficult to structurally observe, more qualitative research methods, such as in-depth interviews and teacher logbooks on how teachers aimed to reach their goals, would have helped to extend our knowledge on what makes (high) goals and positive shifts effective, and for whom.

The use of interviews on gaining information on teachers’ motivation for assigning specific goals for specific students might further have added to our knowledge of the mechanisms that were involved in the goal setting procedure, and might, therefore, further have helped explaining the relations between (high) performance goals, ambition and student achievement. For instance, knowledge of the student characteristics that teachers bear in mind while determining their goals, but also of the teachers’ sense of self-efficacy might have helped to better understand the nature of (high) teacher expectations and goals in natural settings. High-efficacy teachers seem to have fewer negative expectations and are more flexible in adapting them (Tournaki & Podell, 2005). Moreover, high-efficacy teachers are supposed to have higher goal commitment (Locke & Latham, 2002). Understanding the influence of more student and teacher characteristics in the goal setting procedure itself might increase our knowledge of accuracy and positive/negative bias, although it should not be neglected that precisely determining the accuracy of teacher expectations in natural settings is difficult due to the possibility of an omitted variable (Jussim & Harber, 2005; Madon et al., 1997).

A third limitation concerns the way we operationalized the goals. In our study teachers were asked to determine the desired performance levels without involving the students in this goal setting process and without informing the students on what goals

the teachers had set for them. Students only played a passive role in this set-up: teachers wanted them to reach some tacit performance goals. However, as Sadler (1989) argued, students ultimately have to close the gap between the actual and the desired level themselves. As such, students should be aware of what skills and knowledge they are expected to master, what their actual level of performance is and what steps should be taken to reduce the gap. By involving students in monitoring and assessing their own learning activities, students are invited to play an active role in their own learning process and relevant feedback can be given to them (Hattie & Timperley, 2007). By engaging students in the goal-setting process in future studies, students are expected to benefit more from the (high) goals, as both teachers and students take joint responsibility for reaching them.

4.3. Final remarks

In the current study, the relation between high goals, positive changes, and student achievement was investigated. In order to promote well-considered goals a step-by-step procedure was developed, in which teachers were encouraged to think carefully about each student's potential, not only taking into account their earlier performances but also discussing socio-emotional, instructional, and motivational approaches and strategies with their colleagues. This renders the current study an example of research within a broad interpretation of data use (Schildkamp & Kuiper, 2010) in which the full picture of students and their abilities play a role.

References

- Bates, C., & Nettelbeck, T. (2001). Primary school teachers' judgements of reading achievement. *Educational Psychology*, 21(2), 177–187. <http://dx.doi.org/10.1080/01443410020043878>.
- Begeny, J. C., Eckert, T. L., Montarello, S. A., & Storie, M. S. (2008). Teachers' perceptions of students' reading abilities: An examination of the relationship between teachers' judgments and students' performance across a continuum of rating methods. *School Psychology Quarterly*, 23(1), 43–55. <http://dx.doi.org/10.1037/1045-3830.23.1.43>.
- Black, P., & William, D. (2009). Developing the theory of formative assessment. *Educational Assessment, Evaluation and Accountability*, 21(1), 5.
- Brookhart, M. A., Schneeweiss, S., Rothman, K. J., Glynn, R. J., Avorn, J., & Stürmer, T. (2006). Variable selection for propensity score models. *American Journal of Epidemiology*, 163(12), 1149–1156. <http://dx.doi.org/10.1093/aje/kwj149>.
- Caliendo, M., & Kopeinig, S. (2008). Some practical guidance for the implementation of propensity score matching. *Journal of Economic Surveys*, 22(1), 31–72. <http://dx.doi.org/10.1111/j.1467-6419.2007.00527.x>.
- Carlson, D., Borman, G. D., & Robinson, M. (2011). A multistate district-level cluster randomized trial of the impact of data-driven reform on reading and mathematics achievement. *Educational Evaluation and Policy Analysis*, 33(3), 378–398.
- Cizek, G. J., & Bunch, M. B. (2007). *Standard Setting. A guide to establishing and evaluating performance standards on tests*. Thousand Oaks: Sage Publications.
- Datnow, A., Park, V., & Wohlstetter, P. (2007). *Achieving data. how high-performing school systems use data to improve instruction for elementary students*. Los Angeles: University of Southern California.
- de Boer, H., Bosker, R. J., & van der Werf, M. P. C. (2010). Sustainability of teacher expectation bias effects on long-term student performance. *Journal of Educational Psychology*, 102(1), 168–179. <http://dx.doi.org/10.1037/a0017289>.
- Deunk, M. I., van Kuijk, M. F., & Bosker, R. J. (2014). The effect of small group discussion on cutoff scores during standard setting. *Applied Measurement in Education*, 27(2), 77–97. <http://dx.doi.org/10.1080/08957347.2014.880441>.
- Educational Inspectorate (2010). *Opbrengstgericht werken in het basisonderwijs. Een onderzoek naar opbrengstgericht werken bij rekenen-wiskunde in het basisonderwijs*. [Data-driven decision making in primary education. A study on data-driven decision making in mathematics in primary education]. Utrecht, the Netherlands: Inspectorate of Education.
- Expert Group Continuous Learning Progression (2008). *Over de drempels met taal en rekenen [Over the thresholds of language and mathematics]*. Enschede, the Netherlands: Expert group Continuous Learning Progression language and mathematics.
- Feenstra, H., Kleintjes, F., Kamphuis, F., & Krom, R. (2010). *Wetenschappelijke verantwoording begrip lezen groep 3 t/m 6 [Scientific account for the reading comprehension tests grade 1 to 4]*. the Netherlands: CITO: Arnhem.
- Feinberg, A. B., & Shapiro, E. S. (2009). Teacher accuracy: An examination of teacher-based judgments of students' reading with differing achievement levels. *The Journal of Educational Research*, 102(6), 453–462. <http://dx.doi.org/10.3200/JOER.102.6.453-462>.
- Fuchs, L. S., Fuchs, D., & Deno, S. L. (1985). Importance of goal ambitiousness and goal mastery to student achievement. *Exceptional Children*, 52(1), 63–71.
- Fuchs, L. S., Fuchs, D., & Hamlett, C. L. (1989). Effects of alternative goal structures within curriculum-based measurement. *Exceptional Children*, 55(5), 429–438.
- Ganley, C. M., Mingle, L. A., Ryan, A. M., Ryan, K., Vasilyeva, M., & Perry, M. (2013). An examination of stereotype threat effects on girls' mathematics performance. *Developmental Psychology*. Advance online publication. <http://dx.doi.org/10.1037/a0031412>.
- Good, T. L., & Brophy, J. E. (2003). *Looking in classrooms*, 9th ed. Boston: Pearson/Allyn and Bacon.
- Hambleton, R. K., & Pitoniak, M. J. (2006). Setting performance standards. In R. L. Brennan (Ed.), *Educational measurement* (pp. 433–470). Washington DC: American Council on Education.
- Hattie, J., & Timperley, H. (2007). The power of feedback. *Review of Educational Research*, 77(1), 81–112.
- Hattie, J. (2009). *Visible learning. A synthesis of over 800 meta-analyses relating to achievement*. New York: Routledge.
- Hoge, R. D., & Coladarci, T. (1989). Teacher-based judgments of academic achievement: A review of literature. *Review of Educational Research*, 59(3), 297–313.
- Huffman, D., & Kalnin, J. (2003). Collaborative inquiry to make data-based decisions in schools. *Teaching and Teacher Education*, 19(6), 569–580. [http://dx.doi.org/10.1016/S0742-051X\(03\)00054-4](http://dx.doi.org/10.1016/S0742-051X(03)00054-4).
- Janssen, J., Verhelst, N., Engelen, R., & Scheltens, F. (2010). *Wetenschappelijke verantwoording van de toetsen LOVS rekenen-wiskunde voor groep 3 tot en met 8*. [Scientific account of the mathematics assessment for grades 1 to 6]. Arnhem, The Netherlands: Cito.
- Jungbluth, P. (2012). *Leerlingen in de ogen van hun leerkrachten*. [Students observed by their teachers]. Maastricht, The Netherlands: Universiteit Maastricht.
- 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.
- Kluger, A. N., & DeNisi, A. (1996). The effects of feedback interventions on performance: A historical review, a meta-analysis, and a preliminary feedback intervention theory. *Psychological Bulletin*, 119(2), 254–284. <http://dx.doi.org/10.1037/0033-2909.119.2.254>.
- Lai, M. K., & Schildkamp, K. (2013). Data-based decision making: An overview. In K. Schildkamp, M. K. Lai, & L. Earl (Eds.), *Data-based decision making in education. Challenges and opportunities* (pp. 9–21). Dordrecht, The Netherlands: Springer Science+Business Media.
- Ledoux, G., Blok, H., & Boogaard, M. (2009). *Opbrengstgericht werken. over de waarde van meetgestuurd onderwijs*. [Data-driven decision making. About the value of measurement-driven education]. Amsterdam, The Netherlands: SCO-Kohnstamm Instituut.
- Levin, J. A., & Datnow, A. (2012). The principal role in data-driven decision making: Using case-study data to develop multi-mediator models of educational reform. *School Effectiveness and School Improvement*, 23(2), 179–201. <http://dx.doi.org/10.1080/09243453.2011.599394>.
- Linn, R. L., & Baker, E. L. (1995). What do international assessments imply for world-class standards? *Educational Evaluation and Policy Analysis*, 17(4), 405–418.
- Locke, E. A., & Latham, G. P. (1990). *A theory of goal setting and task performance*. Englewood Cliffs, NJ: Prentice Hall.
- Locke, E. A., & Latham, G. P. (2002). Building a practically useful theory of goal setting and task motivation: A 35-year odyssey. *American Psychologist*, 57(9), 705–717. <http://dx.doi.org/10.1037/0003-066X.57.9.705>.
- Locke, E. A., & Latham, G. P. (2006). New directions in goal-setting theory. *Current Directions in Psychological Science*, 15(5), 265–268. <http://dx.doi.org/10.1111/j.1467-8721.2006.00449.x>.
- Madon, S., Jussim, L., & Eccles, J. (1997). In search of the powerful self-fulfilling prophecy. *Journal of Personality and Social Psychology*, 72(4), 791–809.
- Maehr, M. L., & Zusho, A. (2009). Achievement goal theory. The past, present and future. In K. R. Wentzel, & A. Wigfield (Eds.), *Handbook of motivation at school* (pp. 77–104). New York: Routledge.
- Martínez, J. F., Stecher, B., & Borko, H. (2009). Classroom assessment practices, teacher judgments, and student achievement in mathematics: Evidence from the ECLS. *Educational Assessment*, 14(2), 78–102. <http://dx.doi.org/10.1080/10627190903039429>.
- McDonald, L., Flint, A., Rubie-Davies, C., Peterson, E. R., Watson, P., & Garrett, L. (2016). Teaching high-expectation strategies to teachers through an intervention process. *Professional Development in Education*, 42(2), 290–307. <http://dx.doi.org/10.1080/19415257.2014.980009>.
- McKown, C., & Weinstein, R. S. (2008). Teacher expectations, classroom context, and the achievement gap. *Journal of School Psychology*, 46(3), 235–261. <http://dx.doi.org/10.1016/j.jsp.2007.05.001>.
- Means, B., Padilla, C., & Gallagher, L. (2010). *Use of education data at the local level. From accountability to instructional improvement*. Washington, D.C.: U.S. Department of Education.
- Meelissen, M. R. M., & Drent, M. (2008). *TIMMS- Nederland. Trends in leerprestaties in exacte vakken in het basisonderwijs*. [TIMMS-2007 The Netherlands. Trends in student performance levels in mathematics and science education in primary education]. Enschede, The Netherlands: Universiteit Twente.
- Meelissen, M. R. M., Netten, A., Drent, M., Punter, R. A., Droop, M., & Verhoeven, L. (2012). *PIRLS- en TIMSS-2011. Trends in leerprestaties in lezen, rekenen en natuuronderwijs*. [PIRLS and TIMSS 2011. Trends in student performances in

- reading, mathematics and science education]. Enschede, The Netherlands: Universiteit Twente.
- Mitzel, H. C., Lewis, D. M., Patz, R. J., & Green, D. R. (2001). The bookmark procedure: Psychological perspectives. In G. J. Cizek (Ed.), *Setting performance standards. Concepts, methods and perspectives* (pp. 249–281). Mahwah NJ/London: Lawrence Erlbaum.
- Rasbash, J., Browne, W., Healy, M., Cameron, B., & Charlton, C. (2012). *MLwiN version 2.23*. Bristol, England: Centre for Multilevel Modeling.
- Ready, D. D., & Wright, D. L. (2011). Accuracy and inaccuracy in teachers' perceptions of young children's cognitive abilities: The role of child background and classroom context. *American Educational Research Journal*, 48(2), 335–360. <http://dx.doi.org/10.3102/0002831210374874>.
- Ritzema, E. S., Deunk, M. I., & Bosker, R. J. (2016). Differentiation practices in grade 2 and 3: variations in teacher behavior in mathematics and comprehensive reading lessons. *Journal of Classroom Interaction*, 51(2), 49–71.
- Rosenbaum, P. R., & Rubin, D. B. (1985). Constructing a control group using multivariate matched sampling methods that incorporate the propensity score. *The American Statistician*, 39(1), 33–38.
- Rosenthal, R. (1994). Interpersonal expectancy effects: A 30-year perspective. *Current Directions in Psychological Science*, 3(6), 176–179. <http://dx.doi.org/10.1111/1467-8721.ep10770698>.
- Royal Netherlands Academy of Arts, & Sciences (2009). *Rekenonderwijs op de basisschool. Analyse en sleutels tot verbetering. [Mathematics education in primary education. Analysis and keys to improvement]*. Amsterdam, The Netherlands: KNAW.
- Rubie, C. M. (2003). *Expecting the best: Instructional practices, teacher beliefs and student outcomes*. Auckland: University of Auckland.
- Rubie-Davies, C., Hattie, J., & Hamilton, R. (2006). Expecting the best for students: Teacher expectations and academic outcomes. *British Journal of Educational Psychology*, 76(3), 429–444. <http://dx.doi.org/10.1348/000709905X53589>.
- Rubie-Davies, C. M. (2007). Classroom interactions: Exploring the practices of high- and low-expectation teachers. *British Journal of Educational Psychology*, 77(2), 289–306. <http://dx.doi.org/10.1348/000709906X101601>.
- Rubin, D. B. (2007). The design versus the analysis of observational studies for causal effects: Parallels with the design of randomized trials. *Statistics in Medicine*, 26(1), 20–36. <http://dx.doi.org/10.1348/000709905X53589>.
- Südkamp, A., Kaiser, J., & Möller, J. (2012). Accuracy of teachers' judgments of students' academic achievement: A meta-analysis. *Journal of Educational Psychology*, 104(3), 743–762. <http://dx.doi.org/10.1037/a0027627>; 10.1037/a0027627.supp.
- Sadler, D. R. (1989). Formative assessment and the design of instructional systems. *Instructional Science*, 18, 119–144.
- Sammons, P., Hillman, J., & Mortimore, P. (1995). *Key characteristics of effective schools: A review of school effectiveness research*. London: Institute of Education.
- Scheerens, J., & Bosker, R. J. (1997). Size, stability and consistency of school effects. In J. Scheerens, & R. J. Bosker (Eds.), *The foundation of educational effectiveness* (pp. 1997Oxford: Pergamon).
- Schildkamp, K., & Kuiper, W. (2010). Data-informed curriculum reform: Which data, what purposes, and promoting and hindering factors. *Teaching and Teacher Education*, 26(3), 482–496.
- Slavin, R. E. (1990). Achievement effects of ability grouping in secondary schools: A best-evidence synthesis. *Review of Educational Research*, 60(3), 471–499.
- Steiner, P. M., Cook, T. D., Shadish, W. R., & Clark, M. H. (2010). The importance of covariate selection in controlling for selection bias in observational studies. *Psychological Methods*, 15(3), 250–267. <http://dx.doi.org/10.1037/a0018719>.
- Stuart, E. A. (2010). Matching methods for causal inference: A review and a look forward. *Statistical Science*, 25(1), 1–21.
- Tiedemann, J. (2002). Teachers' gender stereotypes as determinants of teacher perceptions in elementary school mathematics. *Educational Studies in Mathematics*, 50(1), 49–62.
- Tournaki, N., & Podell, D. M. (2005). The impact of student characteristics and teacher efficacy on teachers' predictions of student success. *Teaching and Teacher Education*, 21(3), 299–314. <http://dx.doi.org/10.1016/j.tate.2005.01.003>.
- Wayman, J. C., Midgley, S., & Stringfield, S. (2006). Leadership for data-based decision making: Collaborative educator teams. In A. Danzig, K. Borman, B. Jones, & B. Wright (Eds.), *New models of professional development for learner centered leadership* (pp. 189–205). Mahwah, NJ: Erlbaum.
- Wayman, J. C., Jimerson, J. B., & Cho, V. (2012). Organizational considerations in establishing the data-informed district. *School Effectiveness and School Improvement*, 23(2), 159–178. <http://dx.doi.org/10.1080/09243453.2011.652124>; 29.
- Wohlstetter, P., Datnow, A., & Park, V. (2008). Creating a system for data-driven decision-making: Applying the principal-agent framework. *School Effectiveness and School Improvement*, 19(3), 239–259. <http://dx.doi.org/10.1080/09243450802246376>.
- van der Kleij, F. M., & Eggen, T. J. H. M. (2013). Interpretation of the score reports from the computer program LOVS by teachers, internal support teachers and principals. *Studies in Educational Evaluation*, 39(3), 144–152. <http://dx.doi.org/10.1016/j.stueduc.2013.04.002>.
- van der Kleij, F. M., Eggen, T. J. H. M., & Engelen, R. J. H. (2014). Towards valid score reports in the Computer Program LOVS: A redesign study. *Studies in Educational Evaluation*, 43(0), 24–39. <http://dx.doi.org/10.1016/j.stueduc.2014.04.004>.
- Young, V. M. (2006). Teachers' use of data: Loose coupling, agenda setting, and team norms. *American Journal of Education*, 112(4), 521–548.

Evelien, also known as Lienneke, Ritzema is a researcher and lecturer at GION education/research, University of Groningen (The Netherlands). Her research interests include teacher professional development, goal setting, data use, differentiation, and educational effectiveness.

Marjolein Deunk is a researcher and lecturer at GION education/research, University of Groningen (The Netherlands). Her research focuses on differentiation, early childhood education, and language and literacy development in educational settings.

Roel Bosker is full Professor of Education and director of GION education/research, University of Groningen (The Netherlands). His research interests include educational effectiveness, quality care and school self-evaluation, program evaluation, comparative education, and multilevel modeling.

Mechteld van Kuijk is a researcher and lecturer at GION education/research, University of Groningen (The Netherlands). Her research focuses on teacher professional development, reading instruction, cooperative learning, data use, and goal setting.