University of Groningen

Goals, data use, and instruction

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

Published in:
School Effectiveness and School Improvement

DOI:
[10.1080/09243453.2015.1026268](https://doi.org/10.1080/09243453.2015.1026268)

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):

van Kuijk, M. F., Deunk, M. I., Bosker, R. J., & Ritzema, E. S. (2016). Goals, data use, and instruction: the effect of a teacher professional development program on reading achievement. *School Effectiveness and School Improvement*, 27(2), 135-156. <https://doi.org/10.1080/09243453.2015.1026268>

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.

Goals, data use, and instruction: the effect of a teacher professional development program on reading achievement

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

GION education/research, Faculty of Behavioural and Social Sciences, University of Groningen, Groningen, The Netherlands

(Received 21 February 2014; final version received 2 March 2015)

In this paper, we investigated whether student reading comprehension could be improved with help of a teacher Professional Development (PD) program targeting goals, data use, and instruction. The effect of this PD program on 2nd- and 3rd-grade student achievement was examined using a pretest-posttest control group design. Applying propensity score matching, 35 groups in the experimental condition were matched to 35 control groups. Students in the experimental condition ($n = 420$) scored significantly higher on a standardized assessment than the control condition ($n = 399$), the effect size being $d = .37$. No differential effects of the PD program were found in relation to initial reading performance or grade. Different model specifications yielded similar albeit smaller effect sizes ($d = .29$ and $d = .30$). At the end of the program, students in the experimental condition were more than half a year ahead of students in the control condition.

Keywords: program effectiveness; reading achievement; propensity score matching; hierarchical linear modeling

Introduction

As reading constitutes an important basis for learning, working, and living, a common goal of primary schools is to equip students with sufficient reading skills (Kirsch, 2002; Reis, McCoach, Little, Muller, & Kaniskan, 2011; Snow, Burns, & Griffin, 1998). Currently, there are concerns on the reading results of Dutch students in primary school (e.g., Ministerie van Onderwijs, Cultuur en Wetenschap [Ministry of Education], 2008, 2010), particularly the degree to which struggling, poorly performing readers are prepared for later schooling and participation in society (Inspectie van het Onderwijs [Inspectorate of Education], 2007, 2010a). Student assessment results on both national and international reading tests were considered to be insufficient, leading to achievement concerns on the part of policymakers and the general public. For example, although students in The Netherlands perform rather well, comparatively speaking, on the 2011 Progress in International Reading Literacy Study (PIRLS) assessment (targeting fourth-grade reading), the average achievement of the Dutch students is significantly lower than in 2001 (Meelissen et al., 2012). The unsatisfactory student results have been ascribed to various causes, among which the lack of clear performance goals for teachers and schools to aim for in their teaching (Inspectie van het Onderwijs, 2011; Ministerie van Onderwijs, Cultuur en Wetenschap, 2010; Onderwijsraad [Council of Education], 2007). As it is unclear what a sufficient level of reading proficiency entails, it might be difficult for

*Corresponding author. Email: m.f.van.kuijk@rug.nl

teachers to provide very targeted instruction. A related potential cause for these unsatisfactory results pertains to the instruction that teachers provide to their students in the subject area of reading comprehension. In The Netherlands, reading comprehension is a separate subject in the curriculum from the second grade onward. During instruction, the differences between students and their varying needs are not sufficiently met (Bonset & Hoogeveen, 2009; Inspectie van het Onderwijs, 2012). Furthermore, Dutch teachers commonly concentrate on asking questions about the specific text at hand, whereas explicit instruction on the use of reading strategies is frequently missing (Aarnoutse & Weterings, 1995; De Jager, Reezigt, & Creemers, 2002; Droop, Van Elsäcker, & Voeten, 2012; Van Elsäcker, 2002). Yet, text comprehension can be improved by teaching students to use specific strategies when they encounter difficulties in understanding what they are reading. Such strategies are employable to many other texts than only the text at hand; benefitting comprehension in the long run (Andreassen & Braten, 2011; National Reading Panel, 2000; Pressley, 1998). Hence, support has been gained for the view that the desired increase in students' reading performance could be attained by training teachers so that they would become able to set their own performance goals, target differences between students, and provide high-quality reading instruction in which the use of reading strategies is explicitly taught. Our aim was to help Dutch primary school teachers make reading comprehension instruction more goal-oriented, focused, explicit, and better suited to students' needs in order to improve student reading comprehension.

In order to implement new (i.e., more effective) instructional techniques and to change existing routines, teachers need support and guidance (Black & Wiliam, 1998b; Borko, 2004). Teacher Professional Development (PD) programs are a common tool for the implementation of performance improvement efforts (Bishop, Berryman, Wearmouth, & Peter, 2012; Hill, 2007; Organisation for Economic Co-operation and Development [OECD], 2005). Researchers in the area of reading have emphasized the importance of schools and teachers to operate on a "what works" basis, in the context of which scientific evidence is used to improve the instructional practice and student performance (e.g., Armbruster, Lehr, & Osborn, 2010; Collins Block & Lacina, 2009; National Reading Panel, 2000; Pressley, 1998). The call to use scientific evidence as a basis for the adoption of programs and practices not only applies to the subject area of reading but also to the whole field of education (e.g., Slavin, 2008). To help teachers implement more effective practices, we developed a multicomponent teacher PD program. In our PD program, the participating teachers were supported in improving their instructional practice with help of three components: (a) setting standards and performance goals for every student, (b) applying formative assessment and data use, and (c) acquiring relevant instructional skills and (content and curriculum) knowledge in reading comprehension. All three components have shown to be positively related to student performance. As the long-term effects of early acquired literacy skills have been documented in the literature (e.g., Bodovski & Youn, 2011; Snow et al., 1998) and the most striking results of interventions have been found for junior year groups (Heckman, 2006; Mortimore, Sammons, Stoll, Lewis, & Ecob, 1988), our PD intervention program specifically targeted teachers of the second and third grades (student age: approximately 7 to 9 years old).

In this paper, we discuss the set-up of a multicomponent teacher PD program and investigate the effectiveness of this program by comparing reading comprehension levels of students whose teachers participated in the program with those of students whose teachers did not (as the program was developed as a response to the recent performance concerns). Improving student performance via teacher PD programs is difficult (e.g., Desimone, 2009), but, in the review study of Yoon, Duncan, Lee, Scarloss, and Skapley

(2007), the most promising results of teacher professional development on students' reading performance are found for programs that include a focus on how students learn as well as on how to track student learning. For our PD program, we integrated a focus on these elements into one presumably synergetic package – all components being equally important within the PD package. This is why we are interested in *molar causation* (Shadish, Cook, & Campbell, 2002), focusing on the overall relationship between an integral treatment package and its effects. Below, we will elaborate on the PD program's components. After this, additional general information regarding the set-up of the PD program is provided, before continuing to the methods section.

Components underlying the program: theoretical considerations

Component 1: setting standards and performance goals for every student

Our first component pertained to setting goals. Working with goals has generally been proven to be effective for enhancing performance. Setting goals leads to a clearer notion of desired outcomes, and it directs the focus toward the attainment of these desired outcomes (Fuchs, Fuchs, & Deno, 1985; Fuchs, Fuchs, & Hamlett, 1989; Locke & Latham, 1990, 2002; Wiliam, 2010). When working with performance goals, these goals should be defined at a level that challenges teachers and their students, as ambitious goals are associated with higher outcomes (Locke & Latham, 1990, 2002).

As part of the PD program, teachers were asked to set a performance goal for each of their students pertaining to the end of the school year. The goals were formulated by selecting one of five performance categories (labeled below minimum, minimum, basic, proficient, and advanced) in acknowledgement of the differences in students' capabilities. The performance categories had been defined by the participating teachers in an early stage of the PD program with help of a standard setting procedure (e.g., Cizek & Bunch, 2006). During this procedure, teachers studied and discussed reading comprehension assessment items. More information on the standard setting procedure is provided in Deunk, Van Kuijk, and Bosker (2014). A specific feature of the performance categories was that they pertained to test score intervals on the scale of the end-of-the-school year standardized reading comprehension assessments. The performance goals were set in terms of categories which had been linked to these assessments, for example, "I want Billy to attain a score within the proficient category and Julie to attain a score within the advanced category on the standardized reading comprehension assessment". The advantage of setting performance goals in such a manner was that the attainment of these goals would easily be established by conducting the assessment and comparing a student's attained score with the test score interval the teacher had selected for this student. As it was important that these goals were set at an appropriate level for each student given their capabilities, we developed a *multistep procedure* which incorporated performance data analysis and team discussion to help teachers reflect on and reconsider the goals' appropriateness before deciding on its final version. This multistep procedure pertains to the second component of the PD program.

Component 2: applying formative assessment and data use

In order to help the participating teachers set appropriate goals (using the multistep procedure) and, subsequently, to help teachers attain these goals, it was important that they based their instructional decisions on assessment results (e.g., Guskey, 2002). Using

student performance data to adapt one's teaching in order to meet students' needs is known as *formative assessment* (Black & Wiliam, 1998a, 1998b; Herman, Osmundson, & Silver, 2010). In their meta-analysis, Black and Wiliam (1998b) conclude that there is "a body of firm evidence that formative assessment is an essential component of classroom work and that its development can raise ... achievement" (p. 148). Also, other studies have shown that schools and districts applying a "data-driven" way of teaching can result in increased student performance levels (e.g., Carlson, Borman, & Robinson, 2011; Slavin, Cheung, Holmes, Madden, & Chamberlain, 2013). In The Netherlands, the progress of students is monitored with the use of student monitoring systems. Yet, teachers are found to use these systems to a rather limited extent for the purpose of analyzing problems and, subsequently, adapting instruction. Furthermore, teachers who do use the student monitoring systems for these purposes frequently make mistakes in the interpretation of certain outputs (Van der Kleij & Eggen, 2013) and are often unaware of the possibilities for more sophisticated analyses (Ledoux, Blok, & Boogaard, 2009; Meijer & Ledoux, 2011; Schildkamp & Kuiper, 2010; Van der Kleij & Eggen, 2013). Limited use of data by teachers has also been reported in, for example, England (Downey & Kelly, 2013), Flanders (Vanhoof, Verhaeghe, Verhaeghe, Valcke, & Van Petegem, 2010), and South Africa (Archer, Scherman, & Howie, 2013). The limited use of performance data and the student monitoring systems is particularly relevant given the finding that teachers should differentiate more (Inspectie van het Onderwijs, 2012). Differentiation is defined as:

an approach to teaching in which teachers proactively modify curricula, teaching methods, resources, learning activities and student products to address the diverse needs of individual students and small groups of students to maximize the learning opportunity for each student in a classroom. (Tomlinson et al., 2003, p. 121)

When teachers have better insight into the differences between their students in terms of attainment (knowing which students already master a certain skill and which students do not), this will improve the extent to which instruction can suit the needs of different students.

During our PD program, teachers received training in the use of the student monitoring system and in the interpretation of student performance data. An important prerequisite to establish student performance improvement, however, is that teachers adjust their practices accordingly after analyzing the data. This step is not always guaranteed (Goertz, Olah, & Riggan, 2009, as cited in Carlson et al., 2011). Analyzing data "... is not enough to produce gains in achievement. Schools must actually take action to change teaching and learning" (Slavin et al., 2013, p. 390). The third component of our PD program focused on how to take action after analyzing the data.

Component 3: knowledge and instruction for reading comprehension

After setting the performance goals and identifying the progress made toward them, it was important to help the teachers attain their own objectives by ensuring that they were equipped with the most relevant instructional skills and knowledge about reading comprehension instruction. In order to advance students' reading proficiency, it is important that instruction is focused and clear (e.g., Andreassen & Braten, 2011; Verhoeven, 1991). Yet, in The Netherlands, little explicit instruction – particularly in why and how to use reading strategies – is given during lessons in reading

comprehension (Aarnoutse & Weterings, 1995; Van Elsäcker, 2002). This finding is similar to observations of reading comprehension instruction conducted in the United States, Norway, and Belgium (Andreassen & Braten, 2011; Liang & Dole, 2006; Van Keer & Verhaeghe, 2005).

In order to improve the quality of reading comprehension instruction and to improve the degree to which teachers attend to reading strategies, we focused on the reading strategies as discussed in the textbooks that the participating teachers used in their schools. Teachers in The Netherlands are known to follow the curricular textbooks to a very large extent (Meelissen et al., 2012). These textbooks – which are developed by commercial publishers – often contain reading strategies which are not supported by empirical evidence (Droop et al., 2012; Stoeldraijer & Forrer, 2012), and the textbooks are frequently criticized to be “more bulky than necessary, containing a substantial amount of material that has little or nothing to do with learning to read” (Houtveen & Van de Grift, 2012, p. 88). In addition, some textbooks miss certain age-appropriate text genres such as directive texts. During the PD program, we discussed evidence-based reading strategies (e.g., Vernooy, 2005) and age-appropriate reading skills and text genres as identified by the expertise centre for the Dutch language (Expertisecentrum Nederlands, 2010). By doing so, we aimed to make teachers more knowledgeable and skilled in order to be able to make (more) deliberate instructional decisions. During the PD program, teachers were stimulated to critically reflect on the content of the curricular materials and recommendations were provided on how to address shortcomings in these materials.

Furthermore, two evidence-based instructional practices were discussed as part of our PD program to help improve the quality – particularly the explicitness – of reading comprehension instruction. We discussed modeling, which is an instructional technique in which the teacher demonstrates how to apply a reading strategy by thinking aloud and linking the solution to the use of strategies or knowledge that the students already possess (Fisher, Frey, & Lapp, 2008; National Reading Panel, 2000). We also discussed Direct Instruction, which is a teacher-centered model for instruction focused on the content and structure of a lesson (Borman, Hewes, Overman, & Brown, 2003; Muijs & Reynolds, 2011). In the meta-analyses of Swanson and Hoskyn (1998), modeling in combination with Direct Instruction has been identified as an effective instructional model for teaching struggling learners and remediating learning disabilities. This combination entails – among other things – active presentation of information, a clear organization of instruction, and the use of many demonstrations. As part of the PD program, teachers received constructive feedback with respect to their implementation of these practices. For Direct Instruction (a relatively well-known instructional model in The Netherlands, although not well implemented; Inspectie van het Onderwijs, 2010b), we conducted lesson observations at the beginning of the school year to provide teachers with individual feedback on their implementation at that time point; for example, by giving suggestions for improvement when observing partial implementation. We also made slight adaptations to the content of the PD program accordingly (e.g., not discussing this model extensively to teachers who already implemented it well). Anecdotes from the observations were also used to illustrate desired behaviors as part of the program. Feedback on the implementation of modeling was provided during the meeting when this technique was practiced “on the spot” (it was observed rarely at the beginning of the school year). More information on the observations and feedback provided to teachers is discussed in Van Kuijk (2014).

General information on the set-up of the PD program

Second- and third-grade teachers participated in our multicomponent PD program. In addition, the school's principal and the internal support coordinator participated. Principals and internal support coordinators can facilitate teacher change by providing teachers with organizational and emotional *support* (enabling teachers to do their jobs) as well as by providing them a *direction* (by determining a goal for the school or showing enthusiasm on the content of a program, e.g., Fullan, 2001; Harris, 2012; Leithwood, Harris, & Hopkins, 2008). Throughout the school year of 2011–2012, the time investment of the teachers was scheduled for 40 hr, including attending after-school meetings (nine in total; duration 1,5 to 2,5 hr per meeting) and completing homework assignments. Participation was voluntary and free of charge: No incentives (monetary or other) were provided to the participating teachers or schools. More information on the PD program is provided in Van Kuijk (2014).

The current study

In this paper, we investigate the effectiveness of the PD program by comparing reading comprehension levels of students whose teachers participated in the program with those of students whose teachers did not – as we consider the program to be effective when participation in the program has supported teachers to bring about improvements in their students' reading comprehension. In this paper, we also study whether teachers' participation in the program is more effective for certain subgroups of students. As aforementioned, some of the instructional practices (i.e., modeling in combination with Direct instruction) are known to be particularly beneficial to struggling readers, and the most promising results of interventions are found in the junior grades. Therefore, we are interested in the relation between the effect of the teachers' participation in the PD program and students' initial reading performance as well as the relation between the effect of the teachers' participation in the PD program and students' grade. The following research questions are addressed:

- (1) Does teachers' participation in a multicomponent Professional Development program targeting goals, data use, and instruction result in improved student reading comprehension?
- (2) Does the effect of teachers' participation in the Professional Development program on students' reading comprehension depend on students' initial performance or students' grade?

Method

To study the effect of the teacher PD program on reading comprehension achievement, a quasi-experimental pretest-posttest control group design was used.

Participants

Nineteen schools in the northern part of The Netherlands participated in the multicomponent PD program. In total, 33 teachers took part in the program. These teachers taught 33 classes, of which 10 classes were multigrade classrooms which contained both a second- and a third-grade year group. Thus, 43 groups of second- and third-grade students

(containing 20 second-grade and 23 third-grade groups) were taught by the participating teachers. For the remainder of this paper, we will refer to these groups rather than to teachers' classes as the matching procedure focused on the level of the group (i.e., the grade) rather than the class.

Design and construction of the control group

Our study formed part of a larger conglomerate of teacher PD intervention studies. This conglomerate was used to construct a suitable control condition. In total, over 90 Dutch primary schools participated in one (or, in a few cases, two) of five different teacher PD programs offered in the whole series of studies. The other PD programs targeted similar topics such as data use. To promote the participation of the schools, they were given the opportunity to select the PD program of their choice. Each of the intervention studies targeted specific grades in primary school: Our study focused on the second and third grades. In order to construct a control condition for our study, we focused on the second- and third-grade groups from schools that had no intervention in these grades¹, forming a pool of possible control groups. Since our multicomponent PD program was conducted at the group/teacher level, we wanted to match at this level. Ultimately, a number of 80 groups, containing 56 second-grade groups and 24 third-grade groups², were identified as possible controls.

From this pool, we selected those groups that were the most similar to the groups in our experimental condition. Because we intended to take a number of group-level characteristics into account, we applied the propensity score matching approach (Kelcey, 2011; Rosenbaum & Rubin, 1984). In this approach, the relevant group-level variables are combined into one score which is an estimate of the probability that a group participates in the program. This score, with a value between 0 and 1 and estimated through logistic regression, is called the propensity score. The groups that were taught by teachers participating in the PD program were matched to the groups from the pool of possible controls on the basis of this propensity score. The propensity score matching approach ensures that, overall, the variables used in the construction of the propensity score are sufficiently balanced when comparing the experimental and control condition (Rosenbaum, 2009).

In the current study, our pool of possible controls was not extremely large. As we wanted to avoid the risk of selecting poor matches, we applied the caliper matching algorithm.³ In caliper matching, a maximum propensity score distance can be specified between the propensity score of the group in the experimental condition and the propensity score of the group in the control condition (Heinrich, Maffioli, & Vazquez, 2010). By specifying the maximum distance between the matches, the propensity scores of the experimental condition and control condition could not differ too much, reducing the risk of selecting poor matches. A consequence of specifying this caliper can be that not all groups that have participated in the program are provided with a match. The estimation of the effect of the program will then pertain to a subsample of the groups that have participated. In our study, the 0.2 caliper resulted in sufficient matching quality for the experimental and control condition while including as many groups as possible (detailed further).

When using caliper matching, one can match "without replacement" and "with replacement". Using the "without replacement" method, groups from the experimental condition are matched to groups in the control condition on a one-to-one basis. Using the "with replacement" method, a group in the control condition can function as a match for

multiple groups in the experimental condition (Heinrich et al., 2010). Both the 0.2 caliper with replacement and the 0.2 caliper without replacement yielded sufficient matching quality (in terms of balance of the covariates, when comparing the two conditions). In order to select one of these two approaches, we then considered sample size. When applying the 0.2 caliper with replacement, we could include slightly more groups from the experimental condition than when applying the 0.2 caliper without replacement. Yet, this coincided with a reduction in total sample size as fewer students participated in the control condition. This entailed a slight loss of statistical power for the analysis of the effect of the program. The 0.2 caliper without replacement method thus yielded the most preferred results, as it resulted in the largest number of groups and students included in both the experimental and the control condition. As such, the outcomes of the 0.2 caliper without replacement are presented in the results section below. Next, we calculated the effect of the PD program on achievement using the student assessment results of the groups which had been selected via this matching method. As a check for the robustness of our findings, the outcomes of the 0.2 caliper with replacement method are presented in the results section as well.

Instruments and variables

The data collected on the instruments discussed in this paragraph were used for the estimation of the propensity score in order to identify suitable matches. The results on these instruments were also used to investigate the effect of the program on student achievement. An important note pertains to the fact that the matching was conducted at the group level, whereas the analyses on the effect of the program were conducted on the student level (detailed further on, in the analyses section).

Instruments

Reading comprehension assessment. The Cito standardized reading comprehension assessments, (developed by Cito, the Netherlands Institute for Educational Measurement) were used to measure the students' reading comprehension skills. The Cito standardized assessments form part of the assessment system most widely used in The Netherlands; this system – called LOVS – has been employed in approximately 85% of the Dutch primary schools (Inspectie van het Onderwijs, 2010a). In the Cito reading comprehension tests, students are asked to read several texts and answer multiple-choice questions referring to word, sentence, and text levels. The use of these tests has been approved by the Dutch Committee of Tests and Testing (De Commissie Testaangelegenheden Nederland [COTAN]) responsible for the review of tests. Both the validity and reliability of these tests have been considered sufficient: Their reliability is above 0.89 (Cronbach's alpha) for the grades under study (Feenstra, Kleintjes, Kamphuis, & Krom, 2010). The reading comprehension assessments are administered from Grade 1 to Grade 6, and their results are registered on a continuous scale which ranges from –87 to +147 (end of Grade 1 to mid-term Grade 6). The negative symbol (–) for a large part of the assessment scale should not be interpreted as having a negative connotation; –87 is simply the (arbitrary) starting point of this scale.

The pretest results are the results on the Cito reading comprehension assessments which were collected in the year prior to the start of the PD program, that is, in June 2011. These pretest results were used as one of the variables to estimate the propensity score for each group of students. The results were also used as a covariate in the analyses on the

Table 1. Timing of the data collection in reading comprehension and mathematics.

	Timing data collection	End of grade ..	Cito assessment Regular or Additional*
Reading comprehension pretest	June 2011	1 & 2	Regular
Mathematics	June 2011	1 & 2	Regular
Reading comprehension posttest	June 2012	2 & 3	Regular & Additional

*“Regular” refers to assessments which are part of the regular schedule of the standardized assessments in the Cito assessment system. “Additional” refers to the additional multisubject standardized test, developed by Cito, for those assessments which are not regularly conducted in June.

effect of the program on student achievement. The posttest results are the results on the Cito reading comprehension assessments which were collected at the end of the school year in which the PD program was implemented, that is, in June 2012. These posttest results were used to identify the effect of the program on students’ achievement. For most subject areas in the Cito LOVS standardized assessment system, assessments are conducted in each grade in January and in June. However, the reading comprehension assessment has a slightly different timing throughout the grades in primary school: The regular assessments are conducted in June in the first grade, in both January and June in the second grade, and only in January from third grade onward. To obtain an “end-of-the-school-year” (i.e., posttest) result for reading comprehension in the third grade, we used the reading comprehension items of an additional multisubject standardized test (also developed by Cito), which is conducted in June. Its results are registered on the same scale as those of the regular reading comprehension assessments. This information concerning the timing of the reading comprehensions assessments as well as the use of the regular assessment or the additional test is summarized in Table 1.

Mathematics assessment. The Cito standardized mathematics assessments were used in this study as well. When reading, one is actively and deliberately constructing meaning from a text (e.g., Snow, 2002). This can be a cognitively demanding task, which is why we wanted to control for academic ability. Information on students’ IQ was not available, but students’ mathematics results were considered a suitable proxy for general academic ability. The mathematics tests (also part of the Cito LOVS assessment system) are similarly approved by the Dutch Committee of Tests and Testing (COTAN). Both the tests’ validity and reliability have been considered sufficient (Janssen, Verhulst, Engelen, & Scheltens, 2010); their reliability is above 0.91 (Cronbach’s alpha). The mathematics results are registered on a continuous assessment scale which ranges from 0 to 169 (mid-term Grade 1 to mid-term Grade 6). We incorporated students’ prior math performances – as assessed in June 2011 – in the estimation of the propensity score as well as in the analyses on the effect of the program. This information on the timing of the mathematics data collection is also presented in Table 1.

Variables used for the estimation of the propensity score

When estimating a propensity score, one should incorporate variables that are related to the outcome measure and the participation decision (Caliendo & Kopeinig, 2008). Given the small sample in our study, including too many covariates is not recommended (Steiner,

Cook, Shadish, & Clark, 2010). We used seven group-level characteristics to estimate each group's propensity score via logistic regression. These seven variables were:

- (1) grade;
- (2) number of second- or third-grade students in the group;
- (3) percentage of girls in the group;
- (4) the groups' average performance on the pretest;
- (5) the groups' standard deviation of performance on the pretest (indicating the level of heterogeneity in the group);
- (6) the groups' average performance on the mathematics test; and
- (7) the groups' standard deviation of performance on the mathematics test.

We prioritized on including baseline measures of the outcome variable, the aforementioned proxy for academic ability and several other characteristics that could be potentially related to the outcome measure. For example, the variable grade was included as only second- and third-grade classes could participate in the program and we wanted to take the performance differences between these two grades into account. We included the number of students as a variable in the matching procedure as the number of students in classes in the northern (rural parts) of The Netherlands can be quite small (e.g., Veenman, 1995), and some studies have found positive albeit small effects of being taught in classes with smaller student numbers (see Blatchford, 2003). The percentage of girls in a class was included as girls generally outperform boys in reading (e.g., Bond, Dykstra, Clymer, & Summers, 1997). Heterogeneity of the class – both for reading comprehension and mathematics – was included as it has been argued that heterogeneous classes are more difficult to teach than homogeneous classes (discussed in, e.g., Hanushek & Wößmann, 2005). In heterogeneous classes, teachers need to attend to a diversity of instructional needs among students. As a result, students other than those attended by the teacher work a lot without direct support of the teacher (Wilkinson & Hamilton, 2003). For more on the discussion on covariate selection, one is referred to Steiner et al. (2010).

Variables used for the analyses of the effect of the PD program on achievement

After estimating the propensity score and using the outcomes to identify a suitable control condition from a larger pool of possible controls, we compared the results of the experimental condition to those of the control condition in order to study the effect of the program on student achievement. More specifically, we investigated whether students' posttest results (the dependent variable) could be predicted by participation in the PD program (the independent variable) with help of multilevel regression analyses. In these analyses, we controlled for the following covariates: (a) students' sex (using a dummy-coded variable with boys as the reference category), (b) students' grade (using a dummy-coded variable with second-grade as the reference category), (c) students' performance on the pretest (grand-mean centered to facilitate its interpretation), and (d) students' performance on the mathematics assessment (grand-mean centered to facilitate its interpretation). In a preliminary stage of our analyses, we had discovered that the relationship between the pre- and the posttest for reading comprehension was not directly linear but could better be described by including polynomial terms. Including quadratic and cubic transformations of the reading comprehension pretest significantly improved the fit of our models, and these transformations were therefore included. The propensity score itself

was not a significant predictor of student performance and was therefore excluded from the analyses.

Analyses

The *PSmatching* program, an *R* Plugin written as a so-called custom dialog in SPSS, was used to conduct the task of estimating the propensity scores and matching the conditions (Thoemmes, 2012). We investigated the quality of the matching procedure using an independent samples *t*-test, which is a common way to assess the quality of matching procedures. After matching, no significant differences should be found between the conditions on the means of the variables used to estimate the propensity score (Caliendo & Kopeinig, 2008).

In order to assess the effect of the PD program on achievement, a multilevel regression analysis was performed with the help of *MLwiN* software (Rasbash, Browne, Healy, Cameron, & Charlton, 2011), with students (Level 1) nested in groups (Level 2).⁴ In this analysis, it was investigated whether the performance on the posttest could significantly be predicted by participation in the program while controlling for the aforementioned covariates. We studied possible differential effects of teachers' participation in the program on student achievement by including interaction terms in the *MLwiN* analysis.

Results

The quality of the matching results

The results of the propensity score matching approach are discussed first. After using the 0.2 caliper matching algorithm without replacement, it was found that not all groups in the experimental condition could be provided with a match. Of the 43 groups, 35 treated groups were matched to 35 untreated groups. Both conditions contained 19 second-grade groups and 16 third-grade groups. The groups in the experimental condition which could not be matched were characterized by relatively high propensity scores, while the unmatched groups in the pool of possible controls had mostly low propensity scores. In Table 2, we present the descriptive statistics of the variables used to estimate the propensity scores, both before and after matching. The statistical significance of the differences between the experimental and the control condition, tested using the independent samples *t*-test, are – if present – denoted in this table by an asterisk in Column c (in which the experimental and the control condition are compared before matching) and Column f (in which the experimental and the control condition are compared after matching).

As can be seen from the results presented in Table 2, the groups in the experimental condition and in the pool of possible controls differed significantly with respect to three variables (i.e., grade, number of students, and propensity score) prior to matching. After matching, the group-level variables were distributed in a much more balanced way and the differences between the conditions were no longer significant.

The effect of the program on students' reading achievement

After establishing an equivalent experimental and control condition, we assessed the effect of the PD program on student achievement using multilevel regression analyses. The dataset used for these analyses consisted of the results of 819 students in total. The

Table 2. Group characteristics in experimental and control condition before and after matching (means of means).

Variables	Before matching				After matching			
	Experimental (a) (<i>n</i> = 43)		Control (b) (<i>n</i> = 80)		Experimental (d) (<i>n</i> = 35)		Control (e) (<i>n</i> = 35)	
	<i>M</i>		<i>M</i>		<i>M</i>		<i>M</i>	
	<i>(SD)</i>		<i>(SD)</i>		<i>(SD)</i>		<i>(SD)</i>	
Grade	0.54	(0.51)	0.30	(0.46)	0.46	(0.51)	0.46	(0.51)
Number of students	10.49	(6.98)	14.78	(7.24)	12.00	(6.83)	11.40	(5.15)
Percentage of girls	0.50	(0.21)	0.50	(0.14)	0.48	(0.16)	0.48	(0.14)
Average performance on pretest	12.39	(12.99)	8.12	(11.43)	11.11	(13.56)	13.17	(11.27)
<i>SD</i> of performance on pretest	13.42	(6.24)	13.11	(3.38)	13.70	(5.40)	13.24	(3.87)
Average performance on math test	55.95	(12.71)	51.99	(11.42)	54.49	(12.52)	56.11	(11.82)
<i>SD</i> of performance on math test	13.07	(4.84)	13.15	(3.69)	14.05	(4.36)	13.19	(4.37)
Propensity score	0.43	(0.17)	0.30	(0.15)	0.39	(0.17)	0.39	(0.15)

**p* < .05.

Table 3a. Student characteristics in experimental and control condition.

Variables	Experimental <i>n</i> (%)	Control <i>n</i> (%)
Grade 2	231 (55)	210 (53)
Grade 3	189 (45)	189 (47)
Girls	201 (48)	193 (48)
Boys	219 (52)	206 (52)

Table 3b. Descriptives of student performance in experimental and control condition.

Variables	Experimental		Control	
	<i>M</i> (<i>SD</i>)	95% CI	<i>M</i> (<i>SD</i>)	95% CI
Second grade				
Math performance	47.07 (14.71)	45.16; 48.98	47.03 (14.96)	44.99; 49.06
Pretest	2.24 (17.35)	−0.01; 4.49	5.95 (13.91)	4.06; 7.85
Posttest	22.19 (16.02)	19.02; 22.63	20.83 (13.29)	20.11; 24.27
Third grade				
Math performance	64.23 (15.44)	62.02; 66.45	66.54 (15.42)	64.32; 68.75
Pretest	21.33 (14.76)	19.21; 23.44	22.57 (14.54)	20.48; 24.65
Posttest	33.41 (15.55)	31.18; 35.64	31.55 (14.99)	29.39; 33.70

results of 420 students (51%) in the experimental condition were compared to those of 399 students in the control condition. In [Tables 3a](#) and [3b](#), the descriptive statistics of the variables used in the regression analyses are presented. When comparing [Table 2](#) to [Tables 3a](#) and [3b](#), they differ in that the first table refers to the group level and the latter two tables refer to the student level.

The descriptives presented in [Table 3a](#) show that the variables grade and sex were similarly distributed across the conditions. The average performances on the pretests, presented in [Table 3b](#), were above the national averages in both conditions for these timepoints, being $M = -2.75$ and $M = 13.2$. The average performance on the pretest was lower in the experimental group than in the control group, both in second grade and in third grade. As a result of the lower performance results of the students in the experimental condition, we might be at risk of mistaking a *regression-to-the-mean* effect (see Cozby, 2004) for a positive effect of the intervention on the students' achievement. We checked this difference between conditions while taking the multilevel structure of the data (students nested in groups) into account, and found that the differences between the conditions on the pretest were not significant ($t = -1.46$, $p > .05$ and $t = 0.11$, $p > .05$ for the second- and third-grade groups, respectively). The same was done for the mathematics' results. Again, the results were higher than the national average for these timepoints – being $M = 34$ and $M = 56.5$ – and the average performance on the mathematics test was slightly lower in the experimental condition as well. The differences between the conditions were tested and found not to differ significantly ($t = -.28$, $p > .05$ and $t = -.61$, $p > .05$, for the second- and third-grade groups, respectively).

Table 4. Multilevel models predicting achievement in reading comprehension.

Predictors	Models					
	Start model		Main effect model		Interaction model	
	Coeff.	SE	Coeff.	SE	Coeff.	SE
Fixed Part						
Intercept	19.16*	0.92	17.21*	1.03	17.23*	1.17
Grade 3	-3.82*	1.30	-3.74*	1.23	-3.10	1.71
Girls	2.29*	0.75	2.29*	0.74	2.24*	0.75
Math performance	0.17*	0.03	0.16*	0.03	0.17*	0.03
Pretest	0.67*	0.04	0.67*	0.04	0.65*	0.05
Pretest ²	0.00165*	0.00067	0.00155*	0.00067	0.00166*	0.00069
Pretest ³	-0.00006*	0.00001	-0.00006*	0.00001	-0.00006*	0.00001
Participation in PD program			3.78*	1.10	3.85*	1.53
Participation x pretest					0.04	0.05
Participation x grade					-1.27	2.40
Random Part						
Variance at class level	14.03	4.00	11.21	3.53	11.17	3.51
Variance at student level	101.97	5.26	101.74	5.24	101.93	5.25
Deviance	6168.64		6157.17		6156.48	
No. of groups	70		70		70	
No. of students	819		819		819	

* $p < .05$.

Table 4 reports the results of the multilevel analyses. Here, the program's effect on reading achievement was estimated while controlling for the relevant covariates (using the data of 819 students nested in 70 groups). The first model, called the start model, contained the intercept and the covariates at the student and the group level but did not include the variable indicating whether the teacher had participated in the multicomponent PD program. In the second model, called the main effect model, this variable was included. By doing so, we could analyze whether or not this variable added value when predicting student achievement. In the third model, called the interaction model, we investigated differential effects as we added an interaction term between participation in the program and students' pretest performance, as well as an interaction term between participation in the program and students' grade. All models presented in Table 4 contain unstandardized coefficients.

When comparing the start model to the main effect model, one can see that participation in the PD program is related to a significantly higher performance on the posttest. Inclusion of this variable fairly increased the fit of the model: The deviance decreased by 11.47, which is a significant improvement (the critical value in a chi-square distribution with $df = 1$ is 3.84 for $p = .05$, as the models differ in one parameter). Furthermore, inclusion of this variable accounts for 2.6%⁵ of the variance. When comparing the main effect model to the interaction model, the interactions were found to be nonsignificant and the model fit had not improved. In other words, the positive effect of the program on students' achievement was found to apply irrespective of students' initial performance on the pretest, and the PD program was equally effective for second- and third-grade students.

For the effect of the PD program on student achievement, we observed an effect size of $d = .37$ (calculated by dividing its regression coefficient by the square root of the unexplained variance at the student level, using the coefficient in the main effects model⁶), 90% CI [.20; .55]. According to Cohen's interpretation (1988)⁷, a value of $d = .37$ is a small to medium effect.

Checks for robustness

To check the robustness of our results, we used different model specifications. First, we investigated whether participation in the program remained a significant predictor of achievement when separately modeling the outlying residuals at both the student and the group levels. Outlying cases might have an "undue high influence on the results of the statistical analysis" (Snijders & Bosker, 1999, p. 128), and we wanted to ensure that the positive effect of participation on achievement – as identified in the current study – was not caused by influential outliers. In identifying the outliers, we made use of z scores.⁸ In the outlier model we fitted, participation in the PD program was a significant positive predictor of student achievement with an effect size of $d = .29$, 90% CI [.13; .45].

The second approach we used to check the robustness of our results was application of the 0.2 caliper with replacement matching method. When matching with replacement, the control groups are given larger weights in the analyses of the effect of the program as they are matched to multiple groups in the experimental condition. Using the 0.2 caliper with replacement method, 20 second-grade and 20 third-grade groups in the experimental condition were matched to 14 second-grade and 15 third-grade groups in the control condition. Five previously excluded groups in the experimental condition could now be provided with a match; thus, more experimental groups participated in the estimation of the effect of the program. The differences between the experimental and the control condition for the variables used to construct the propensity score were nonsignificant (tested by applying the independent samples t -test). When using the assessment results of the students in the groups selected via this matching method, we found a similar effect size of $d = .30$, 90% CI [.12; .48] for the effect of teachers' participation in the PD program on the students' reading achievements.

In summary, we used two alternative model specifications to check for the robustness of our original results, and these alternative methods yielded similar positive outcomes for the effect of the PD program on student achievement – with slightly smaller but fairly comparable effect sizes.

Conclusion and discussion

In this paper, we investigated the effect of teachers' participation in a multicomponent teacher Professional Development (PD) program on student reading comprehension. This PD program had been developed as a response to the recent performance concerns in The Netherlands. Given the importance of the early acquired literacy skills, we specifically targeted second- and third-grade teachers in this improvement effort. The PD program was designed to foster student reading comprehension through teachers' application of a multicomponent package on goals, data use, and instruction. Using the propensity score matching approach to construct an equivalent control condition from a larger pool of possible controls, we found that students in the experimental condition performed significantly better than those in the control condition, with an effect size of $d = .37$; 90% CI [.20; .55]. We checked for the robustness of these results using different model

specifications, and found similar albeit smaller effect sizes for the effect of the PD program on student achievement ($d = .29$ and $d = .30$, respectively). According to Cohen's interpretation (1988), these effect sizes can be interpreted as small to medium effects. One has to bear in mind that these results compare to approximately 60% of a grade equivalent. At the end of the program, students in the experimental condition were more than half a year ahead of students in the control condition. If one takes into account that the experimental condition and control condition already performed above the national average at the start of the program, it is considered a promising finding that students in the experimental condition outperform students in the control condition at the end of the school year. That being said, we consider a replication of this study including schools and teachers serving a population of lower achieving readers to be a valuable supplement to the current study, particularly since the program was designed following performance concerns and we did not find differential effects of the program. In our study, all students – irrespective of whether they were initially low- or high-achieving students or whether they were in second or third grade – appeared to have profited equally from their teachers' participation in the PD program. Even though several aspects of the PD program were known to be particularly effective for low-achieving readers, it might be that the PD program *as a whole* was not specific enough to improve the results of struggling readers and to narrow the achievement gap in reading comprehension. Perhaps more intensive didactical practices in addition to the ones used in our program are necessary. A replication study with three conditions (one control condition, and two experimental conditions with one condition pertaining to the original version of the program and one condition pertaining to a modified version of the program focusing on struggling readers) would shed more light on this finding.

Several limitations to the current study should be considered. In order to facilitate participation, schools could choose whether or not they wanted to participate in our PD program – they were not randomly assigned to the experimental or the control condition. In order to account for differences between these conditions, we used the propensity score matching approach, but the number of variables used to estimate the propensity score as well as the number of groups in the pool of possible controls was relatively small. We originally wanted to include teacher characteristics (such as teachers' attitude toward data use) in order to incorporate variables which might have been of influence in the decision to participate in the program. Yet, the degree of non-response to a teacher questionnaire, one of the instruments used in the larger series of intervention studies, prohibited us from doing so. Nonetheless, all schools in the series of intervention studies participated because they wanted to improve their education through participation in PD programs targeting similar topics. Therefore, relatively similar schools and teachers are considered to have taken part in both the experimental and the control condition. Nevertheless, a replication of this study in which schools are randomly assigned to conditions would allow for stronger statements on the causation of the program's effects, that is, there would be no threat of omitted variable bias. Thus, random assignment to conditions should be included in the design of our aforementioned “three-condition” replication study.

Another consideration pertains to the intensity of the program and the fact that the participants knew that they were trained by researchers. We cannot refute that the positive effect of our PD program on student achievement could have been caused by a *Hawthorne* effect (Shadish et al., 2002). In such a case, the participants would have improved their behavior simply because of the knowledge that they were being studied, and not because of the content of our program. However, all schools and teachers

participating in the series of intervention studies – including teachers in the control condition – were aware of the fact that students' results were measured throughout the entire school for evaluation purposes. This makes the Hawthorne effect a less likely explanation for the results of our study. In addition, to attain reading performance improvement as we did in our study, teachers are assumed to have provided high-quality instruction. For this, we consider the content of our program to have been essential. Not all teacher professional development programs targeting reading succeed in significantly improving student results (see the overview of studies in Yoon et al., 2007). As aforementioned, the most promising results of PD programs on reading contained a focus on content and data (Yoon et al., 2007), and a successful stimulation of reading performance might require PD programs combining these elements. This line of reasoning seems to be confirmed by the multicomponent *Success for All* program (Slavin et al., 1996) targeting the prevention of and early intervention in reading difficulties. In this program, among other things, teachers use a prescribed curriculum in order to provide high-quality instruction, and students' progress is frequently monitored. After 3 years of continuously implementing this *Success for All* program, effect sizes between $d = .21$ and $d = .36$ were found in kindergarten to Grade 2 across various reading performance measures (Borman et al., 2007).

The last consideration pertains to the fact that we, at this time, are left with the question *how* teachers' participation in the program resulted in teacher change that subsequently resulted in improved student achievement.

In the current paper, we studied the effect of a multicomponent teacher PD program which was aimed at improving student reading comprehension. The significant higher reading results of students in the experimental condition – which were found to be robust across different model specifications – lend support for the conclusion that the program was successful in attaining its aim. The results in our study reinforce the findings in other studies where positive effects on student performance are found for programs that combine a subject-specific focus with data use and goal setting.

What practical implications can be derived from the results of this study? Researchers and practitioners in the field of teacher professional development are recommended to move on along this path of working with goals, data use, and subject-specific instruction. Governments and policymakers can facilitate participation in such professional development programs through different means, being (a) legislation, (b) financial stimuli, and (c) communication and moral appeal (Leune, 2001). Such governmental support, in whichever shape or form, can stimulate in-service teacher professionalization. Moreover, we recommend integrating the content of three components into initial teacher training programs. The training of various skills – such as the setting of performance goals, the use of the student monitoring systems, and providing instruction via modeling and Direct Instruction – as well as a stronger knowledge base for teaching reading comprehension are considered recommended for this setting. Both efforts (targeting current teachers and future teachers) are considered important. Results from the current study underline the potential gains that can be made in teaching reading comprehension.

Acknowledgements

The authors would like to express their gratitude to Henk Guldmond, who prepared the software needed to conduct the propensity score matching.

Disclosure statement

No potential conflict of interest was reported by the author(s).

Notes

1. Our PD program was the only one explicitly including the school principal and internal support coordinator (in addition to the participating teachers). This is why, for the other PD programs as part of the larger series of intervention studies, we did not expect any “contamination” of the second- and third-grade teachers via the school principal or internal support coordinator.
2. This relatively small number of third-grade classes can be explained by the use of the multi-subject test at the end of third grade to obtain an end-of-the-school-year measurement of student reading comprehension: The use of this test is detailed in the instruments section of the current paper. A precondition for inclusion of a third-grade class in the pool of possible controls was that this additional multisubject test had been administered.
3. There are different matching algorithms, and each matching algorithm has its own advantages and disadvantages – for more information, see Caliendo and Kopeinig (2008) and Heinrich, Maffioli, and Vazquez (2010). “It should be clear that there is no ‘winner’ for all situations [...]”. Pragmatically, it seems sensible to try a number of approaches. Should they give similar results, the choice may be unimportant” (Caliendo & Kopeinig, 2008, pp. 44–45). We tested several options and found robust results in terms of balance among the variables used to estimate the propensity score.
4. Even though the group-level is not the same as the classroom level (as some groups were part of a multigrade class), we refer to the group/teacher level for sake of simplicity as the results were found to be similar when higher levels of nesting were accounted for. More complex hierarchical models, including students nested in groups, nested in classrooms/teachers, nested in schools, were fitted to account for this structure of the data. The results of these models were almost the same as those of the model reported in this paper; the effect of the program remained significant and had a similar size.
5. The total unexplained variance of the start model is 116, and the total unexplained variance of the main effect model is 112.95 (see Table 4). $R^2 = 1 - [112.95/116] = 0.0263$.
6. This is an application of Cohen’s (1988) formula of $d = \{\bar{x}(exp) - \bar{x}(control)\} / \sigma$ to a multilevel setting, for which we are interested in the variation within groups (i.e., Level 1).
7. Cohen (1988) provides the following guideline for the interpretation of effect sizes: $d = 0.2$ is considered to be a small effect, $d = 0.5$ a medium effect, and $d = 0.8$ a large effect.
8. At the student level, outliers were defined as values with standardized scores lower than $z = -3.29$ or larger than $z = 3.29$ (Tabachnick & Fidell, 2001). Three students could be flagged as having an outlying score after fitting the model, which was caused by an extremely high result on the posttest. For the standardized residuals at the classroom level, we used a stricter z -score criterion, as outlying cases at the classroom level may influence the model more substantially than outlying cases at the pupil level (Rasbash, Steele, Browne, & Goldstein, 2012). In total, four classes with a standardized residual below $z = -2$ or above $z = 2$ were modeled separately. In addition to using this z -score approach to outlier identification, we also checked the influence of outliers with help of a method proposed by Tukey (1977) which makes use of P25 (the first quartile), P75 (the third quartile), and the Inter-Quartile Range (IQR): Here, outliers have values below $P25 - 1.5 \times IQR$ and above $P75 + 1.5 \times IQR$. This approach yielded a larger effect size for the program, with $d = .40$, 90% CI [.23; .58].

Notes on contributors

Mechteld van Kuijk is a researcher and lecturer at GION education/research, University of Groningen (The Netherlands). Her research interests include teacher professional development, data use, reading comprehension, goal setting, and educational effectiveness.

Marjolein Deunk is a researcher and lecturer at GION education/research, University of Groningen (The Netherlands). Her research focuses on language, reading, and literacy development in educational contexts, data use, differentiation, and small primary schools.

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

Evelien, also known as Lieneke, Ritzema is a PhD student at GION education/research, University of Groningen (The Netherlands). Her research focuses on teacher professional development, goal setting, data use, differentiation, and mathematics.

References

- Aarnoutse, C., & Weterings, A. (1995). Onderwijs in begrijpend lezen [Education in reading comprehension]. *Pedagogische Studiën*, 72, 82–101.
- Andreassen, R., & Braten, I. (2011). Implementation and effects of explicit reading comprehension instruction in fifth-grade classrooms. *Learning and Instruction*, 21, 520–537.
- Archer, E., Scherman, V., & Howie, S. (2013). Approaches to effective data use: Does one size fit all? In K. Schildkamp, M. K. Lai, & L. Earl (Eds.), *Data-based decision making in education* (pp. 91–112). Dordrecht: Springer.
- Armbruster, B. B., Lehr, F., & Osborn, J. (2010). *Put reading first. The research building blocks for teaching children to read: Kindergarten through grade 3*. Jessup, MD: National Institute for Literacy.
- Bishop, A. R., Berryman, M. A., Wearmouth, J. B., & Peter, M. (2012). Developing an effective education reform model for indigenous and other minoritized students. *School Effectiveness and School Improvement*, 23, 49–70.
- Black, P., & Wiliam, D. (1998a). Assessment and classroom learning. *Assessment in Education: Principles, Policy & Practice*, 5, 7–63.
- Black, P., & Wiliam, D. (1998b). Inside the black box. *Phi Delta Kappan*, 80, 139–148.
- Blatchford, B. (2003). *The class size debate: Is small better?* Maidenhead: Open University Press.
- Bodovski, K., & Youn, M. J. (2011). The long term effects of early acquired skills and behaviors on young children's achievement in literacy and mathematics. *Journal of Early Childhood Research*, 9, 4–19.
- Bond, G., Dykstra, R., Clymer, T., & Summers, E. G. (1997). The cooperative research program in first-grade reading instruction. *Reading Research Quarterly*, 32, 348–427.
- Bonset, H., & Hoogeveen, M. (2009). *Lezen in het basisonderwijs* [Reading in primary school]. Enschede: SLO.
- Borko, H. (2004). Professional development and teacher learning: Mapping the terrain. *Educational Researcher*, 33(8), 3–15.
- Borman, G. D., Hewes, G. M., Overman, L. T., & Brown, S. (2003). Comprehensive school reform and achievement: A meta-analysis. *Review of Educational Research*, 73, 125–230.
- Borman, G. D., Slavin, R. E., Cheung, A. C. K., Chamberlain, A. M., Madden, N. A., & Chambers, B. (2007). Final reading outcomes of the national randomized field trial of Success for All. *American Educational Research Journal*, 44, 701–731.
- Caliendo, M., & Kopeinig, S. (2008). Some practical guidance for the implementation of propensity score matching. *Journal of Economic Surveys*, 22, 31–72.
- 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, 378–398.
- Cizek, G. J., & Bunch, M. B. (2006). *Standard setting: A guide to establishing and evaluating performance standards on tests*. Thousand Oaks, CA: SAGE.
- Cohen, J. (1988). *Statistical power analysis for the behavioral sciences*. Hillsdale, NJ: Lawrence Erlbaum.
- Collins Block, C., & Lacina, J. (2009). Comprehension instruction in kindergarten through grade three. In S. E. Israel & G. G. Duffy (Eds.), *Handbook of research on reading comprehension* (pp. 494–509). New York, NY: Routledge.
- Cozby, P. C. (2004). *Methods in behavioral research* (8th ed.). New York, NY: McGraw Hill.
- De Jager, B., Reezigt, G. J., & Creemers, B. P. M. (2002). The effects of teacher training on new instructional behaviour in reading comprehension. *Teaching and Teacher Education*, 18, 831–842.

- Desimone, L. M. (2009). Improving impact studies of teachers' professional development: Toward better conceptualizations and measures. *Educational Researcher*, 38, 181–199.
- 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, 77–97.
- Downey, C., & Kelly, A. (2013). Professional attitudes to the use of data in England. In K. Schildkamp, M. K. Lai, & L. Earl (Eds.), *Data-based decision making in education* (pp. 69–89). Dordrecht: Springer.
- Droop, M., Van Elsäcker, W., & Voeten, R. (2012). *Effecten van de BLIKSEM-aanpak in groep 5 & 6 van het basisonderwijs* [Effects of the BLIKSEM approach in third and fourth grade in primary school]. Nijmegen: Expertisecentrum Nederlands.
- Expertisecentrum Nederlands. (2010). *Van leerlijnen naar tussendoelen: Begrijpend lezen* [From learning paths to short-term goals: Reading comprehension]. Nijmegen: Author.
- Feenstra, H., Kleintjes, F., Kamphuis, F., & Krom, R. (2010). *Wetenschappelijke verantwoording begrijpend lezen groep 3 t/m 6* [Scientific account for the reading comprehension tests Grades 1 to 4]. Arnhem: Cito.
- Fisher, D., Frey, N., & Lapp, D. (2008). Shared readings: Modeling comprehension, vocabulary, text structures, and text features for older readers. *The Reading Teacher*, 61, 548–556.
- Fuchs, L. S., Fuchs, D., & Deno, S. L. (1985). Importance of goal ambitiousness and goal mastery to student achievement. *Exceptional Children*, 52, 63–71.
- Fuchs, L. S., Fuchs, D., & Hamlett, C. L. (1989). Effects of alternative goal structures within curriculum-based measurement. *Exceptional Children*, 55, 429–438.
- Fullan, M. (2001). *The new meaning of educational change*. New York, NY: Teachers College Press.
- Guskey, T. R. (2002). Professional development and teacher change. *Teachers and Teaching: Theory and Practice*, 8, 381–391.
- Hanushek, E. A., & Wößmann, L. (2005). Does educational tracking affect performance and inequality? Differences-in-differences evidence across countries. *The Economic Journal*, 116, 63–76.
- Harris, D. M. (2012). Varying teacher expectations and standards curriculum differentiation in the age of standards-based reform. *Education and Urban Society*, 44, 128–150.
- Heckman, J. J. (2006). Skill formation and the economics of investing in disadvantaged children. *Science*, 312, 1900–1902.
- Heinrich, C., Maffioli, A., & Vazquez, G. (2010). *A primer for applying propensity score matching* (Technical Notes No. IDB-TN-161). Retrieved from <http://www.iadb.org/en/publications/publication-detail,7101.html?id=8240%20>
- Herman, J. L., Osmundson, E., & Silver, D. (2010). *Capturing quality in formative assessment practice: Measurement challenges* (CRESST Report No. 770). Los Angeles, CA: CRESST.
- Hill, H. C. (2007). Learning in the teaching workforce. *The Future of Children*, 17(1), 111–127.
- Houtveen, A. A. M., & Van de Grift, W. (2012). Improving reading achievements of struggling learners. *School Effectiveness and School Improvement*, 23, 71–93.
- Inspectie van het Onderwijs. (2007). *De staat van het onderwijs: Jaarverslag 2005–2006* [The state of education: Year report 2005–2006]. Utrecht: Author.
- Inspectie van het Onderwijs. (2010a). *De staat van het onderwijs: Onderwijsverslag 2008–2009* [The state of education: Year report 2008–2009]. Utrecht: Author.
- Inspectie van het Onderwijs. (2010b). *Opbrengstgericht werken in het basisonderwijs* [Data-driven teaching in elementary schools]. Utrecht: Author.
- Inspectie van het Onderwijs. (2011). *De staat van het onderwijs: Onderwijsverslag 2009–2010* [The state of education: Year report 2009–2010]. Utrecht: Author.
- Inspectie van het Onderwijs. (2012). *De staat van het onderwijs: Onderwijsverslag 2010–2011* [The state of education: Year report 2010–2011]. Utrecht: Author.
- Janssen, J., Verhulst, N., Engelen, R., & Scheltens, R. (2010). *Wetenschappelijke verantwoording van de toetsen LOVS rekenen-wiskunde voor groep 3 t/m groep 8* [Scientific account for the mathematics tests Grades 1 to 6]. Arnhem: Cito.
- Kelcey, B. (2011). Assessing the effects of teachers' reading knowledge on students' achievement using multilevel propensity score stratification. *Educational Evaluation and Policy Analysis*, 33, 458–482.
- Kirsch, I. S. (2002). *Reading for change: Performance and engagement across countries. Results from PISA 2000*. Paris: OECD.

- Ledoux, G., Blok, H., & Boogaard, M. (2009). *Opbrengstgericht werken: Over de waarde van meetgestuurd werken* [Data use: The value of data-driven teaching]. Amsterdam: SCO Kohnstamm Instituut.
- Leithwood, K., Harris, A., & Hopkins, D. (2008). Seven strong claims about successful school leadership. *School Leadership and Management*, 28, 27–42.
- Leune, J. M. G. (2001). *Onderwijs in verandering: Reflecties op een dynamische sector* [Education in change: Reflections on a dynamic sector]. Groningen: Wolters-Noordhoff.
- Liang, L. A., & Dole, J. A. (2006). Help with teaching reading comprehension: Comprehension instructional frameworks. *The Reading Teacher*, 59, 742–753.
- Locke, E. A., & Latham, G. P. (1990). *A theory of goal setting & 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, 705–717.
- 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 performance in reading, mathematics and science]. Nijmegen: Radboud Universiteit Nijmegen.
- Meijer, J., & Ledoux, G. (2011). *Gebruikersvriendelijke leerlingvolgsystemen in het primair onderwijs* [User-friendliness of monitoring and evaluation systems in primary education]. Amsterdam: SCO Kohnstamm Instituut.
- Ministerie van Onderwijs, Cultuur en Wetenschap. (2008). *Beleidsreactie doorlopende leerlijnen taal en rekenen* [Policy response continuous learning trajectories language arts and mathematics]. Den Haag: Author.
- Ministerie van Onderwijs, Cultuur en Wetenschap. (2010). *Invoering referentieniveaus taal en rekenen* [Implementation performance standards language arts and mathematics]. Den Haag: Author.
- Mortimore, P., Sammons, P., Stoll, L., Lewis, D., & Ecob, R. (1988). *School matters: The junior years*. Wells: Open Books.
- Muijs, D., & Reynolds, D. (2011). *Effective teaching: Evidence and practice*. London: SAGE.
- National Reading Panel. (2000). *Teaching children to read: An evidence-based assessment of the scientific research literature on reading and its implications for reading instruction*. Washington, DC: National Institute of Child Health and Human Development.
- Onderwijsraad. (2007). *Versteviging van kennis in het onderwijs II* [Reinforcement of knowledge in education II]. Utrecht: Author.
- Organisation for Economic Co-operation and Development. (2005). *Teachers matter: Attracting, developing and retaining effective teachers*. Paris: Author.
- Pressley, M. (1998). *Comprehension strategies instruction. literacy for all: Issues in teaching and learning*. New York, NY: Guilford Press.
- Rasbash, J., Browne, W., Healy, M., Cameron, B., & Charlton, C. (2011). MLwiN version 2.23 [Computer software]. Bristol: Centre for Multilevel Modeling.
- Rasbash, J., Steele, F., Browne, W., & Goldstein, H. (2012). *A user's guide to MLwiN: Version 2.26*. Bristol: Centre for Multilevel Modelling.
- Reis, S. M., McCoach, D. B., Little, C. A., Muller, L. M., & Kaniskan, R. B. (2011). The effects of differentiated instruction and enrichment pedagogy on reading achievement in five elementary schools. *American Educational Research Journal*, 48, 462–501.
- Rosenbaum, P. R. (2009). *Design of observational studies*. New York, NY: Springer.
- Rosenbaum, P. R., & Rubin, D. B. (1984). Reducing bias in observational studies using subclassification on the propensity score. *Journal of the American Statistical Association*, 79, 516–524.
- Schildkamp, K., & Kuiper, W. (2010). Data-informed curriculum reform: Which data, what purposes, and promoting and hindering factors. *Teaching and Teacher Education*, 26, 482–496.
- Shadish, W. R., Cook, T. D., & Campbell, D. T. (2002). *Experimental and quasi-experimental designs for generalized causal inference*. Boston, MA: Houghton Mifflin Company.
- Slavin, R. E. (2008). Perspectives on evidence-based research in education – What works? issues in synthesizing educational program evaluations. *Educational Researcher*, 37, 5–14.
- Slavin, R. E., Cheung, A., Holmes, G. C., Madden, N. A., & Chamberlain, A. (2013). Effects of a data-driven district reform model on state assessment outcomes. *American Educational Research Journal*, 50, 371–396.

- Slavin, R. E., Madden, N. A., Dolan, L. J., Wasik, B. A., Ross, S., Smith, L., & Dianda, M. (1996). Success for all: A summary of research. *Journal of Education for Students Placed at Risk (JESPAR)*, 1, 41–76.
- Snijders, T. A. B., & Bosker, R. J. (1999). *Multilevel analysis: An introduction to basic and advanced multilevel modeling*. London: SAGE.
- Snow, C. E. (2002). *Reading for understanding: Toward an R&D program in reading comprehension*. Santa Monica, CA: Rand Corporation.
- Snow, C. E., Burns, M. S., & Griffin, P. (1998). *Preventing reading difficulties in young children*. Washington, DC: National Academies Press.
- Steiner, P. M., Cook, T. D., Shadish, W. R., & Clark, M. (2010). The importance of covariate selection in controlling for selection bias in observational studies. *Psychological Methods*, 15, 250–267.
- Stoeldraijer, J., & Forrer, M. (2012). *Effectiever en efficiënter werken aan begrijpend lezen* [Working on reading comprehension skills more effectively and efficiently]. Utrecht: PO-raad, School aan Zet.
- Swanson, H. L., & Hoskyn, M. (1998). Experimental intervention research on students with learning disabilities: A meta-analysis of treatment outcomes. *Review of Educational Research*, 68, 277–321.
- Tabachnick, B. G., & Fidell, L. S. (2001). *Using multivariate statistics*. Boston, MA: Allyn and Bacon.
- Thoemmes, F. (2012). *Propensity score matching in SPSS*. Tübingen: University of Tübingen, Center for Educational Science and Psychology.
- Tomlinson, C. A., Brighton, C., Hertberg, H., Callahan, C. M., Moon, T. R., Brimijoin, K.,... Reynolds, T. (2003). Differentiating instruction in response to student readiness, interest, and learning profile in academically diverse classrooms: A review of literature. *Journal for the Education of the Gifted*, 27, 119–145.
- Tukey, J. W. (1977). *Exploratory data analysis*. Reading, MA: Addison-Wesley.
- 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, 144–152.
- Van Elsäcker, W. (2002). *Development of reading comprehension: The engagement perspective* (Doctoral dissertation). Retrieved from <http://repository.ubn.ru.nl/handle/2066/19128>
- Van Keer, H., & Verhaeghe, J. P. (2005). Comparing two teacher development programs for innovating reading comprehension instruction with regard to teachers' experiences and student outcomes. *Teaching and Teacher Education*, 21, 543–562.
- Van Kuijk, M. F. (2014). *Raising the bar for reading comprehension: The effects of a teacher professional development program targeting goals, data use, and instruction* (Doctoral dissertation). Retrieved from <http://irs.ub.rug.nl/ppn/372456987>
- Vanhoof, J., Verhaeghe, G., Verhaeghe, J. P., Valcke, M., & Van Petegem, P. (2010). The influence of competences and support on school performance feedback use. *Educational Studies*, 37, 141–154.
- Veenman, S. (1995). Cognitive and noncognitive effects of multigrade and multi-age classes: A best-evidence synthesis. *Review of Educational Research*, 65, 319–381.
- Verhoeven, L. (1991). Begrijpend lezen in de aanvangsfase: Rol van coherentie, inferentie en anafora [Reading comprehension in the early years: The role of coherence, inference and anafora]. In P. Reitsma, & M. Walraven (Eds.), *Instructie in begrijpend lezen* (pp. 113–127). Delft: Eburon.
- Vernooy, K. (2005). *Elke leerling een competente lezer!* [Each pupil a competent reader!]. Amersfoort: CPS.
- Yoon, K. S., Duncan, T., Lee, S. W. Y., Scarloss, B., & Skapley, K. (2007). *Reviewing the evidence on how teacher professional development affects student achievement*. Washington, DC: National Center for Education Evaluation and Regional Assistance.
- William, D. (2010). The role of formative assessment in effective learning environments. In H. Dumont, D. Istance, & F. Benavides (Eds.), *The nature of learning: Using research to inspire practice* (pp. 135–159). Retrieved from http://www.oecd-ilibrary.org/education/the-nature-of-learning_9789264086487-en
- Wilkinson, I. A. G., & Hamilton, R. J. (2003). Learning to read in composite (multigrade) classes in New Zealand: Teachers make the difference. *Teaching and Teacher Education*, 19, 221–235.