Ms.: Special Issue JEM

Assessing and Validating Effects of a Data-Based Decision-Making Intervention on Student Growth for Mathematics and Spelling

Corresponding author
Prof. Dr. Ir. Jean-Paul Fox
Department of Research Methodology, Measurement and Data Analysis
University of Twente
Postbus 217
7500 AE Enschede
The Netherlands
Phone: +3153 489 3326
j.p.fox@utwente.nl

Co-authors
Dr. Trynke Keuning
Department of Research Methodology, Measurement and Data Analysis
University of Twente
Postbus 217
7500 AE Enschede
The Netherlands
Phone: +3153 489 5564
t.keuning@utwente.nl

Dr. Marieke van Geel
Department of Research Methodology, Measurement and Data Analysis
University of Twente
Postbus 217
7500 AE Enschede
The Netherlands
Phone: +3153 489 5564
E-mail: marieke.vangeel@utwente.nl

Prof.Dr. Adrie J. Visscher
Department of Research Methodology, Measurement and Data Analysis
University of Twente
Postbus 217
7500 AE Enschede
The Netherlands
Phone: +3153 489 5564
a.j.visscher@utwente.nl

Assessing and Validating Effects of a Data-Based Decision-Making Intervention on Student Growth for Mathematics and Spelling

Date May, 2019

Abstract

Data-based decision making (DBDM) is presumed to improve student performance in elementary schools in all subjects. The majority of studies in which DBDM effects have been evaluated has focused on mathematics.

A hierarchical multiple single-subject design was used to measure effects of a two-year training, in which entire school teams learned how to implement and sustain DBDM, in 39 elementary schools. In a multilevel modeling approach, student achievement in mathematics and spelling was analyzed to broaden our understanding of the effects of DBDM interventions. Student achievement data covering the period from August 2010 to July 2014 were retrieved from schools' student monitoring systems. Student performance on standardized tests was scored on a vertical ability scale per subject, for grades one to six. To investigate intervention effects, linear mixed effect analysis was conducted. Findings revealed a positive intervention effect for both mathematics and spelling. Furthermore, low-SES students and low-SES schools benefitted most from the intervention for mathematics.

Keywords: data-based decision making, linear mixed models, multiple single-subject design

Assessing and Validating Effects of a Data-Based Decision-Making Intervention on Student Growth for Mathematics and Spelling

Throughout the last decade, policy makers around the globe have increasingly emphasized the use of data in education to enhance student achievement (Orland, 2015; Schildkamp, Ehren, & Lai, 2012). As a result, the number of reform initiatives aimed at promoting 'data-based decision making' (DBDM), or 'data-driven decision making' (DDDM) have increased rapidly (e.g. Boudett, City, & Murnane, 2005; Carlson, Borman, & Robinson, 2011; Love, Stiles, Mundry, & DiRanna, 2008; Ritzema, 2015; Schildkamp, Poortman, & Handelzalts, 2015; Slavin, Cheung, Holmes, Madden, & Chamberlain, 2012). In these initiatives teachers are encouraged to use data such as student achievement scores on standardized tests and or curriculum-based tests to monitor students' progress, to identify students' needs and to adapt instruction based on this information (Lai & Schildkamp, 2013; Mandinach, 2012). The idea of using student achievement data for evaluating student progress, providing tailor-made instruction, and developing strategies for maximizing performance in order to positively influence student outcomes, seems straightforward. However, the number of large-scale studies into the effects of DBDM on student outcomes is limited and the studies available have mainly focused on the effects of DBDM interventions on students' mathematics outcomes (e.g. Henderson, Petrosiono, Guckenburg, & Hamilton, 2007; Ritzema, 2015; van Geel, Keuning, Visscher, & Fox, 2016) rather than on reading comprehension, vocabulary or spelling. In order to broaden our understanding of the effects of DBDM on student outcomes, research into DBDM-effects on multiple subjects is necessary. A few studies have examined the effects of data-use on reading (e.g. Carlson et al., 2011; Konstantopoulos, Miller, & van der Ploeg, 2013; Quint, Sepanik, & Smith, 2008), but to our knowledge, studies into the effects of DBDM on students' spelling outcomes are non-existent.

The University of Twente developed a DBDM intervention in which entire elementary school teams were systematically introduced to DBDM and trained. Teachers learned how to analyze data, set goals, and choose appropriate instructional strategies based on the data, and, finally to alter instruction in the classroom accordingly.

In 2011, the DBDM intervention showed promising results on mathematics outcomes for a first group of 53 elementary schools. In a group of 7,500 students, a statistically significant positive improvement of student achievement, of approximately one extra month of schooling was achieved during two intervention years, was found. Furthermore, the results suggested that the intervention had been particularly effective at improving the performance of students in low socio-economic status schools (van Geel et al., 2016).

The current DBDM intervention study is similar to the former one, but a different set of elementary schools are considered and an additional topic is considered. Therefore, the aim of the current study is to investigate whether previously found intervention effects can be generalized to a larger population covering multiple topics (i.e., math and spelling), but also to validate the findings of the first study. The internal and external validity of the quasi-experimental design, which is used in both studies, are improved through a novel multilevel design. It is shown that by fulfilling several strict conditions (Kratochwill et al., 2010), causal inferences can be made about the measured intervention effects at the level of schools. It is claimed that under these conditions, the results of the current study can be used to validate the results of the former study.

A Quasi-Experimental Study Design For Evaluating School-Wide Interventions

All participating schools followed the two-year DBDM intervention, where the intervention was applied school wide. However, it was not possible to randomly assign schools to a control condition. Schools made commitments to participate in this project, and most schools preferred to be assigned to the treatment condition because the intervention program promised to improve the student performances. In some cases schools had doubts about the efficacy of the program and wanted to be assigned to the control group. This self-selection of schools to their conditions threatened the external validity, since participating schools would be different from nonparticipating schools (Ji, DuBois, Flay, & Brechling, 2008). Therefore, schools were recruited without a randomization process to obtain an adequate sample size in numbers and representativeness, where each school was assigned to the treatment. In a completely randomized recruitment process, the number of schools being randomized is typically small, which will also not ensure equivalence between treatment and control conditions (Flay & Collins, 2005)The non-randomization procedure to select schools was chosen to maximize the likelihood of recruiting schools. As a result, to collect the data a novel multiple single-subject design was used, where the schools were measured repeatedly over time. In this quasi-experimental design, previous achievements of participating schools were used as a baseline, and school improvements were measured during the intervention and compared to the baseline. Although the schools are the unit of analysis to assess the effects of the DBDM intervention, large numbers of students were selected in each school to ensure statistical power in the study, and to ensure that each school was accurately characterized. Students across grade years from each school were repeatedly measured before and during the intervention to obtain accurate school measurements. The scores of students across grade years were measured on a vertical scale using tests from the student monitoring system (e.g. Vlug, 1997). An improvement in scoring on this vertical scale is considered to be achievement growth, which is represented by a change in scale scores. The tests have been developed through item response theory (IRT) techniques, and it has been shown that they lead to accurate and reliable performance scores (Janssen & Hickendorff, 2010). Furthermore, the spelling and mathematics tests have been rated good by the Dutch Committee on Testing (COTAN) (De Wijs, Kamphuis, Kleintjes, & Tomessen, 2010; Janssen, Verhelst, Engelen, & Scheltens, 2010). By averaging these student's performance scores across grade years, accurate and precise school measurements were obtained, which were robust against extreme scoring students.

Furthermore, to obtain reliable and accurate school-specific intervention effects, the information from all schools were pooled by combining the results from the multiple single-school studies. Thus, in contrast to the typical small sample sizes, which are often used in single-subject studies, the statistical power and reliability is greatly enlarged by using large numbers of students per school to measure school-specific effects, and by pooling the information from all schools. Not all students were repeatedly measured over time during the four years of data collection, since each year new students entered first grade whereas other students left primary education after grade six (see also Figure 2). However, the multilevel modeling approach can handle an unbalanced (data) design, in which students differ in their number of measurements.

From a multilevel modeling perspective, it is known that the students are nested in schools, and the students can be considered lower-level units, where schools are the higher-level units. The schools (level-2 units) were repeatedly measured in this study, where the student population changed over study years. The design extends the hierarchical single-subject design of Van den Noortgate and Onghena (2003) and Jenson, Clark, Kircher, and Kristjansson (2007). In their approach, the level-1 units are repeatedly measured, and the level-2 unit is defined to pool the results.

Theoretical Framework

In the following section, first the rationale underlying the assumption that DBDM positively influences student outcomes is explained. Second, we explain characteristics of effective DBDM interventions aimed at improving student outcomes. Next, after a brief description of the DBDM intervention, we briefly present the results of the previous study into the effects of the DBDM intervention on mathematics. Finally, the hypotheses for this study will be presented.

The Link between DBDM and Student Outcomes

Ikemoto and Marsh (2007) use the following definition of DBDM: “teachers, principals, and administrators systematically collecting and analyzing data to guide a range of decisions to help improve the success of students and schools” (Ikemoto & Marsh, 2007, p. 108). The data is supposed to inform educators, for example, for making deliberate instructional decisions, choosing a new curriculum, or for selecting a proper professional development intervention for their district. This data can encompass anything, from student results on benchmark assessments, student daily work, curriculum-based tests, homework, to classroom observations (Supovitz, 2012). In general, it is assumed that DBDM has a positive influence on student outcomes (Turner & Coburn, 2012). The rationale behind this assumption can be found in the scientific evidence concerning the power of feedback. Data can provide feedback to boards or districts, schools, and teachers on how students, teachers, and schools perform in comparison to the national average, whether student progress is adequate, and on how students perform on subject matter content elements. Although feedback is not a panacea (Kingston & Nash, 2011), the positive performance improving effects of using feedback and formative assessment have been shown in several reviews and meta-analyses (Black &

Wiliam, 1998; Fuchs & Fuchs, 1986; Hattie, 2009; Hattie & Timperley, 2007; Kluger & DeNisi, 1996; Van der Kleij, Vermeulen, Schildkamp, & Eggen, 2015).

Over the past ten years, a substantial number of studies has investigated DBDM. Several special issues regarding data-use reflect the growing interest in DBDM (e.g. Coburn & Turner, 2012; Mandinach & Gummer, 2015; Schildkamp et al., 2012; Schildkamp & Lai, 2013b; Turner & Coburn, 2012). The majority of studies has focused on the effects of DBDM initiatives on teachers' attitudes, knowledge and behavior. Fewer studies have aimed at investigating student outcomes, the final criterion for DBDM-effectiveness. These studies, in which the effect of DBDM on student achievement was studied, mainly focused on mathematics and/or reading outcomes (e.g. Carlson et al., 2011; Konstantopoulos et al., 2013; Ritzema, 2015).

To our knowledge, this is the first study where effects of DBDM on spelling are investigated. Spelling is important for both writing and reading (Graham & Santangelo, 2014). Especially students from a low socio-economic backgrounds run a higher risk to develop impaired spelling, consequently influencing their writing and reading skills (Graham et al., 2008). As studies into the effects of DBMD on mathematics achievement have shown that DBDM was especially beneficial for low-SES students, the intervention may yield similar benefits for spelling.

DBDM is not subject-specific, educators are stimulated to implement DBDM across all subjects. However, effects of DBDM on student performance may vary across subjects. To broaden our understanding of the connection between DBDM and student outcomes, interventions applied to a variety of subjects should be examined.

The Challenge of Impacting Student Outcomes

In Figure 1 (Keuning, Van Geel, Fox, & Visscher, 2016), the four components of DBDM are shown. It was expected that DBDM interventions that include all four DBDM components in a coherent and consistent way would have the largest impact on student achievement. Data-based decision making starts with analyzing data, but it encompasses much more. As Kaufman (2014) states: "While identifying and analyzing data lays the groundwork for impactful improvements to student learning, the resulting actions and progress monitoring will ultimately determine the efficacy of DDDM efforts" (p. 341). Many DBDM interventions mainly focus on the first component of DBDM, and it was found that this does not necessarily lead to changes in teacher classroom practices, not to mention changes in student outcomes (Ikemoto & Marsh, 2007; Marsh, Pane, & Hamilton, 2006; Oláh, Lawrence, & Riggan, 2010). It seems therefore essential that, in order for DBDM interventions to be meaningful and effective, the interventions include *all* DBDM components. From a logical point-of-view, the first component from Figure 1, analyzing and evaluating data, is only meaningful if it is part of the entire DBDM cycle. If data-analysis is not combined with goal-setting and the adaptation of instruction it is unlikely that student achievement improves. Based on the insights gained from the analysis of data, SMART (Specific, Measurable, Attainable, Relevant, Time-bound) and challenging goals should be set. Next, strategies for accomplishing these goals have to be chosen and, finally, the chosen strategy should be executed. Since DBDM is ideally carried out in a systematic approach, data are also supposed to be used for monitoring and evaluating the effects of the implemented strategy, so that the extent to which goals have been achieved can be evaluated, and new data-informed decisions can be made.

[INSERT FIGURE 1 ABOUT HERE]

A second characteristics a DBDM-intervention, as shown in Figure 1, is that the process of DBDM (ideally) takes place at the board, school and class level. However, research has repeatedly shown that, of the malleable factors within a school, teachers influence student outcomes most (Darling-Hammond, 2000; Hattie, 2009; Kaufman et al., 2014; Nye, Konstantopoulos, & Hedges, 2004). Many DBDM initiatives have not involved the teacher level sufficiently. Sometimes, interventions were only implemented at the district level and teachers were unaware of their participation in a DBDM reform (e.g. McCaffrey & Hamilton, 2007). In other cases, interventions were aimed at only training the school leader (e.g. Slavin et al., 2012) or a subset of motivated teachers (e.g. Schildkamp & Poortman, 2015). This is often done under the assumption that a school leader or a small group of teachers will 'spread the word' throughout the entire school, but examples show that this expectation is not always fulfilled. In the so-called data-team procedure (Schildkamp & Poortman, 2015), a group of teachers and a school leader collaboratively learn how to use data to deal with problems faced within the school. In one study, data-team results were received skeptically by other staff members who had not been involved from the outset in data team activities (Schildkamp & Poortman, 2015). Slavin et al. (2012) argued that "helping school leaders to understand student data is helpful but in itself does not produce educationally important gains in achievement" (p. 390).

In sum, we assume that to positively influence student achievement, a DBDM intervention should pay attention to the class/teacher level and, at that level, to the whole DBDM package, instead of only a few DBDM elements. We assume that student outcomes will improve once a DBDM intervention meets these two prerequisites, regardless of the subject the intervention focuses on. The University of Twente developed a DBDM-intervention in line with these recommendations, which will be described now.

The DBDM intervention

Student monitoring system. In the Netherlands over 90% of schools have a student monitoring system (SMS). Such a system includes a coherent set of tests for the longitudinal assessment of students' achievement throughout all grades of elementary education. These test, that are developed by the Central Institute of Test Development are usually taken twice a year in January and in July by all students (Kamphuis & Moelands, 2000). The tests are available for all core subjects (mathematics, reading, spelling and vocabulary) and can best be described as interim benchmark assessments. Teachers enter students' test results into their SMS-software. Thereafter, graphs and tables representing various aspects of student performance can be retrieved from the system. The SMS-software also allows for comparison between student scores and national benchmarks. The tests are clearly designed for monitoring student achievement progress and analyzing patterns in achievement across students and grades and are therefore generally not perceived as 'high-stakes' tests (Kamphuis & Moelands, 2000). These data from the student monitoring system were the starting point for the DBDM-intervention.

Outline of the intervention. The DBDM intervention consisted of a two-year training course for entire Dutch elementary school teams (all teachers as well as the members of the management team such as the school leader and deputy director), aimed at implementing and sustaining DBDM in the whole school organization by systematically following the DBDM cycle as shown in Figure 1.

Table 1 provides an overview of the first and second intervention year meetings.

[INSERT TABLE 1 ABOUT HERE]

The first year of the intervention included seven team meetings aimed at developing DBDM knowledge and skills. The first four meetings were primarily aimed at DBDM-related knowledge and skills: analyzing and interpreting test score data from the student monitoring system, diagnosing learning needs, setting performance goals, and developing instructional plans. Prior to the fifth meeting, teachers had executed the instructional plans in the classroom and, based on students' curriculum-based tests, classwork, homework and classroom observations, they had adjusted those plans, if necessary. At the time of the fifth meeting, the DBDM cycle had been completed for the first time and student achievement data were then discussed in a team meeting. During this meeting, teachers shared their experiences with effective and ineffective classroom practices. Meeting six focused on collaboration among team members by preparing them for observing each other's lessons; either to learn from the colleague they visited, or to provide him/her with feedback. In the last meeting of the school year, the DBDM cycle was completed for the second time and student results and classroom practices were evaluated again. Furthermore, teachers made an instructional plan for the next school year (and for the teacher(s) of that year) and also provided class information to the new teacher. In addition to the seven meetings, teachers were provided with feedback by the external trainer on both the way they had analyzed and interpreted data, as well as on the quality of their instructional plans. Furthermore, teachers were provided with a feedback report concerning their teaching skills as judged by their students.

The second intervention year was aimed at deepening, sustaining and broadening DBDM within the school and included five meetings, in which new subjects were introduced (optional for schools). The DBDM cycle was completed again twice that year. Furthermore, a coaching session was included in this second school year, in which the DBDM trainer observed teachers' classroom instruction and provided them with feedback. This coaching component was added to the intervention, to support teachers in the final step of the DBDM-cycle: ‘executing strategies for goal accomplishment’.

Integration of features of effective teacher professional development. Aside from the two criteria for DBDM-interventions discussed in the previous section (the inclusion of *all DBDM-components* and *the **intensive** involvement of teachers*), in the development of the intervention, the features of effective teacher professional development were also taken into account (Desimone, 2009; Van Veen, Zwart, & Meirink, 2011). We describe these features and the method of integrating them into the intervention, in the following paragraphs.

A clear *link between newly learned knowledge and skills and the practice of schools* is considered essential (Timperley, 2008; Van Veen et al., 2011). Therefore, when learning how to analyze data, teachers applied what they had learned about data on their own students. Furthermore, in the instructional plans teachers learned to develop, they set goals and formulated instructional strategies to achieve these goals for their own classes.

During the meetings teachers engaged in *active learning*, they for example discussed their data analysis results in small groups or investigated the alignment of standardized test components and the curriculum.

Since it takes time to learn and change, *duration* is an important feature of effective professional development in two ways: the number of contact hours and the time span over which the TPD activity is spread (Birman, Desimone, Porter, & Garet, 2000; Desimone, 2009; Garet, Porter, Desimone, Birman, & Yoon, 2001). Due to the many other obligations teachers face in their work, they should be provided with sufficient time to master the learning goals (Timperley, 2008; Van Veen et al., 2011). Hence, the DBDM intervention in this study persisted for two years. The first year included seven contact meetings (each one lasted approximately four hours) and participants were encouraged to apply what they had learned in practice, for example, by carrying out data analyses, developing instructional plans, and finally, adapting their instruction (Timperley, 2008; Van Veen et al., 2011).

Finally, *collective participation* (e.g. as a school team) is often positively associated with active participation in professional development activities. Garet et al. (2001), Lumpe (2007), and Van Veen, et al. (2011) as well as Timperley (2008) argue that interaction and collaboration between colleagues is important for mastering and implementing an innovation. Therefore, the entire school team participated in the DBDM intervention.

By taking into account the features of effective TPD, by engaging all teachers in the training, and by paying attention to all elements of the DBDM cycle, we expected the DBDM intervention to influence teaching quality and student outcomes positively.

Results of the Previous Study on Mathematic Outcomes

In 2011, a first group of 53 elementary schools participated in the DBDM intervention (van Geel et al., 2016). Results of this study indicated that a DBDM intervention in which whole school teams are actively involved and in which attention is paid to all DBDM components can improve students' mathematics outcomes.

Using linear mixed models, an average positive intervention effect of approximately 1.40 ability score points (*S.E.* = .31) was found, indicating an average effect of almost one extra month of schooling during the two intervention years. This statistically significant effect was found for a group of approximately 7,500 students. The random part of the multilevel model showed that this intervention effect varied significantly across schools, whereas the correlation between the random intercept and the random intervention effect of *r* = .84 suggested that the intervention effect was smaller for schools with high initial achievement. Moreover, the intervention effect was larger for schools with high proportions of low-SES students, compared to schools with few low-SES students. At the student level, a significant positive intervention-effect for low-SES students compared to medium-SES students was found.

The Current Study

The previous study had provided evidence that the DBDM-intervention improved student outcomes, however, that study only focused on mathematics. To test whether the DBDM intervention would also show similar positive effects on student outcomes for other subjects (i.e. spelling) and to strengthen the generalizability of the findings for mathematics, we conducted a conceptual replication (Makel & Plucker, 2014; Schmidt, 2009) of the 2011 study.

In August 2012, a new group of 43 schools started a DBDM intervention similar to the DBDM intervention described in the previous. The same aim, implementing and sustaining DBDM in the whole school organization, was pursued by delivering a training in working with the (entire) DBDM cycle. In this intervention, the same content was taught to the participants and the training included the same sequence of meetings as the previous study. One extra classroom coaching session was added to the program in the second intervention year, to ensure that all teachers would be provided with feedback on the execution of DBDM within the classroom. However, the training was not led by the same DBDM trainers. These trainers had also been appointed by the University of Twente for this project and the implementation of the training was supervised by the first author of this paper, who was not directly involved in working with the schools.

The major difference between this project and the 2011 study was that it was up to the participating schools to decide whether they wanted to start the intervention with the implementation of DBDM for mathematics or spelling. In the second year, as was done in the 2011-study, schools could again choose to continue with DBDM for the subject that had been chosen in the first year or to broaden the scope to another subject. This enabled us to investigate the effects of the intervention on both mathematics and spelling.

Hypotheses

In line with the 2011-study, it was expected that, as a result of the intervention, student achievement growth would increase for both mathematics (*hypothesis 1A*) and spelling (*hypothesis 1B*).

Next to these main intervention effects, it was expected that school-specific intervention effects would differ across schools (*hypothesis 2).* It was expected that the chosen trajectory would influence the intervention-effects. Since the duration of the intervention influences the effectiveness of implementation (Timperley, 2008, Van Veen et al., 2011) we expected that schools that implemented DBDM for two years for the same subject would benefit most from the intervention for that specific subject compared to schools that choose to broaden the scope from spelling to for example mathematics in the second intervention year.(*hypothesis 3*).

Moreover, we assumed that school-SES would partly explain differences in intervention-effects between schools: in schools with a high proportion of low-SES students, the intervention-effect was expected to be higher *(hypothesis 4)*. These schools, on average, score lower than schools with a high-SES student population (Carlson et al., 2011; Inspectie van het Onderwijs, 2012) and in the Netherlands teachers are more likely to underestimate the potential of students from a low-SES background (CBS Statline, 2018; Inspectie van het Onderwijs, 2018). Since the intervention was aimed at developing ambitious goal-setting by teachers and improving the educational achievement of *all* students, the intervention-effect was expected to be higher in low-SES schools. At the student level, the intervention effect was expected to be highest for low-SES students for the same reason (*hypothesis 5*).

Based on large-scale studies such as TIMMS (Mullis, Martin, Foy, & Arora, 2012), that showed that student characteristics 'gender' and 'age' were correlated with student outcomes, these variables were also taken into account in our analyses. Finally, at the school level the background variables 'school size' and 'urbanization' were included.

Methodology

Data for this study were gathered from 39 participating elementary (K-6) schools in the Netherlands from August 2012 until July 2014. Student achievement data covering the period of August 2010 until July 2014 were retrieved from the student monitoring systems of the schools. In this section, the sample and method of data collection are described first, after which a description of the data analysis methods will be presented.

Sample

In August 2012, 42 schools started with the DBDM intervention. Two schools dropped-out during the two intervention-years because of a mismatch between intervention content and school challenges at the time. One school was founded in the year 2011, therefore no data was gathered in the period before the intervention, as the school had not yet existed. This school was therefore excluded from this sample. The final sample included 39 participating schools. Characteristics of these schools are presented in Table 2.

[INSERT TABLE 2 ABOUT HERE]

The average school size was 238.4 students (range = 79-530) and was categorized into small (a maximum of 150 students), medium (151 – 350 students), and large (more than 350 students). Seven schools were situated in the four biggest cities in the Netherlands, and thus located in urban areas, 15 school were located in suburban areas (i.e. middle tot large size Dutch towns) and 17 schools were located in more rural areas.

In Dutch educational policy, the level of parents' education is used as proxy for socioeconomic status (SES). Three SES categories can be distinguished (Inspectie van het Onderwijs, 2013), students with 'low SES' (maximum parental educational level: primary education, or special needs education), students with 'medium' SES (maximum parental educational level: lowest level of secondary vocational education, or not more than two years of secondary education), and students with 'high SES' (parental education is at least medium level of secondary vocational education). Since the median educational level in the Netherlands is tertiary vocational education, the students labeled as 'medium SES' cannot be regarded as 'average SES'; both categories medium SES and low SES are below the national average. Dutch schools receive additional funding based on these SES-categories.

In this study, school-SES was based on the percentage of students with low, medium, and high SES, where the proportion of low-SES students were considered to be comparable to four times the proportion of medium-SES students. This is based on the additional funding schools receive for low-SES and medium-SES students, where a medium-SES and a low-SES student count as 0.3 and 1.2 additional student, respectively.

For instance, a school with 15% low and/or medium SES students can be composited as (15-4x)% medium-SES students and x% low-SES students, where x is greater equal zero and less than 15/4. According to this rule, three SES categories were distinguished: high-SES schools (schools with less than (15-4x)% medium-SES students and x% low-SES students), low-SES schools (schools with more than (18+x)% low-SES students and (82-4x)% medium SES students), and medium-SES schools (schools not classified as low or high SES).

During the first year of the intervention, schools chose one main intervention subject (mathematics, spelling, vocabulary or reading), to focus on. After one year, they were given the option to add a second subject or to continue working with only the main subject. Half a year later (so after one and a half year of the intervention) schools were again given the choice to work on a new subject. This approach resulted in different intervention trajectories across schools, such as “spelling-spelling-mathematics”, indicating that this school focused on spelling during the first year and the beginning of the second year, and added mathematics in the second half of the second year (‘more than one year spelling’ in Table 2). Five schools did not use the DBDM intervention for mathematics at all and were therefore excluded from analyses involving the development of mathematics achievement. For spelling, 11 schools did not use the DBDM intervention for spelling and were removed from the analysis of the development in spelling results.

Next, students with only one measurement point were removed from the sample, since they did not contribute to measuring performance growth. For mathematics 494 students were removed as there was only one measurement available; for spelling 482 students were removed. The majority of these students were in grade 6 when data gathering started, meaning that after the first measurement point these students left the school; this is also illustrated in Figure 2. This resulted in a sample of 8,023 unique students for mathematics (40,711 observations) and 6,610 unique students for spelling (34,861 observations). Table 3 presents the characteristics of these students.

[INSERT TABLE 3 ABOUT HERE]

Measures and Data Collection

Results on the mathematics test and the spelling test from the schools' student monitoring system were used to measure student achievement growth. The test results can be converted into an ongoing ´vertical ability scale for per subject that enables the monitoring of student progress over grades and school years. The student performance for grades one to six (students aged six to twelve years old) on these standardized tests for *mathematics* and *spelling* were used in this study. As can be seen in Figure 2 there are eleven test scores during a students' school career (two measurements per grade for grade years one to five, and one for grade six). The score range of math was 0 to 168 and spelling 66 to 197. Over the course of the two years preceding the intervention and the two intervention years, a maximum of eight measurements was observed. Not all students participated in the study for the entire period, which led to an incomplete design with varying number of measurements across students. For instance, for a student who started in grade 3 in school year 2013-2014 only two measurements were observed.

In addition to students' ability scores, we collected student level data concerning gender, socioeconomic status (SES) category (high, medium, low) and age. Age was converted on the basis of average age in months at the time of the test, next age was centered around the mean. As such the age variable is indicating how many months younger or older a student was than expected based on the time of the test. At the school level, data were collected on school size, school SES and urbanization (see Table 2).

[INSERT FIGURE 2 ABOUT HERE]

Multiple Single-Subject Design

Our sample did not allow us to treat any schools as controls. However, multiple measurements prior to the intervention (baseline measurements) and multiple measurements during the intervention (treatment phase) were made to collect valuable information about school performance during and prior to the intervention. Hence, by comparing the performance of the schools from the period prior to the intervention to the period during the intervention, schools served as their own controls.

In this study, schools were repeatedly measured before and during the intervention using measurements of performance of their students. As can be seen in Figure 2, mathematics- and spelling performance was measured repeatedly over time, both before the intervention period (the control phase) and during the intervention period (the treatment phase). The student population of each school changed over time, which did not make it possible to consider differences in performances of each student before and during the intervention. Per student only 2 to 8 sequential measurements were observed leading to an unbalanced design at the student level. At a higher level a balanced design was given, where each school in the study was measured twice a year for a period of two years before and two years during the intervention. Therefore, a single-subject design applied to each school for the eight-sequential school-average measurements. Combining the single-subject designs across schools led to a multiple single-subject design for all schools in the study. This feature of the study design made it possible to measure a general intervention effect for the schools in the study and school-specific intervention effects. Each school-specific measurement was based on several student measurements, but the repeated school measurements were based on different groups of students. As a result, a hierarchical multiple single-subject design was used to measure the intervention effects, where students were nested in schools and schools and students were measured over time. The hierarchical aspect of the study design addressed that students were measured a different number of times and were nested in schools.

Using this so-called hierarchical multiple single-subject design and fulfilling some strict conditions, it was possible to make causal inferences in studies without a control group (Kratochwill et al., 2010). When following the guidelines of Kratochwill et al. (2010), four criteria are set to meet the evidence standards. First, the intervention was designed to improve student achievement, where the start and implementation of the intervention was completely under control of the researchers. Second, standardized tests were used to measure student performance, and they were evaluated to have a high standard of inter-rater reliability (Kamphuis & Moelands, 2000). The tests were constructed by Cito, a large test developer institute in the Netherlands. Their tests are well-known for their good psychometric properties, and the used tests had a reliability above .90 (Janssen et al., 2010, Table 5). Third, multiple attempts were made to assess the intervention effect, although this was done across schools. Obviously, it was not possible to implement multiple baselines within each school, but in this study multiple treatment effects were measured across schools. The intervention was implemented at the same time across schools, which means that the baseline was not set at three different time points. However, this last restriction typically applies to a single subject, where we have considered multiple single subjects. Fourth, in this study each school was followed for a period of two years before and during the intervention. A substantial number of measurements were made within each school for each period of two years.

Handley, Lyles, McCulloch, and Cattamanchi (2018), argued that in a real-world setting the quasi-experimental design has its merits, specifically when randomization is not possible. However, care should be taken in actions to improve the internal and external validity. Therefeore, the evidence-based aspect of the intervention can be further supported by discussing the balance between internal (e., the degree to which errors are minimized) and external validity (e.g., the degree to which results can be generalized to the population). First, with respect to the internal validity of the study, the repeated measurements at the school level, at the pre-intervention period, did not show a typical pattern which could indicate effects of threats as instrumentation, maturation, and statistical regression. Each school measurement is constructed as the average of independent student measurements, and its measurement error variance is represented by the average measurement error variance of the student scores divided by the number of students. As result, the school measurements have a high precision, since a large number of students within each school were used to construct the school measurements. This also diminished the chance on extreme school measurements due to sampling or measurement errors. The repeated school measurements before and during the intervention contributed to a more reliable and accurate estimate of the intervention effects, and increased the internal validity.

Second, by using information of the change in performance of other schools, it was possible to increase the reliability and accuracy of the estimation of a school's intervention effect. Third, the average intervention effect was based on multiple school-specific intervention effects and were therefore also robust against bias from event effects.

The schools were measured eight times, and an interruption was expected half way, where the intervention started. This might have opted for an interrupted time-series design, where the serial correlation between school measurements is directly modeled and the object is to identify a change in the trend. However, a straightforward interrupted time-series approach was not possible. In general, eight correlated measurements (i.e., four before and four during the intervention) are not considered sufficient to identify a significant change in the trend (Shadish, Cook, & Campbell, 2002). Therefore, a joint modeling approach is needed across the schools in the study to obtain sufficient information about a general change in the trend. Furthermore, the change in performance at the school level can only be interpreted conditionally on measured change in student performance. Therefore, a hierarchical design is needed, which also includes the change in performance of repeatedly measured students.

Model

In a hierarchical modeling approach, the growth in student performance was modeled conditionally on the change in school performance. The multilevel modeling approach adopted the unbalanced design at the level of students, and random effects were used to model the dependencies among students in the same grade. However, due to this unbalanced design, student performance could not be modeled per grade level. This would have led to a huge number of random effect parameters and a complex missing data problem, since many students were not measured at each grade level. This problem was avoided by measuring average student performance in grade class 1-3 (middle level) and grade class 4-6 (higher level) and a baseline level, which was the first test occasion in the third grade. Therefore, it was not possible to study differences in intervention effects across grades.

At the student level, three student random effects, representing individual differences (i.e., at mid-grade 1, grade 1-3, and grade 4-6) from the school-average scores, were used to model the growth in performance. The correlation between the student random effects were assumed to capture the serial correlation in repeated measurements of each student.

Next to the three student random effects, two random effects at the school level were introduced to model the variation in performance across schools before the intervention and during the intervention. This random effect during the intervention represents the school-specific intervention effect, and represents a homogenous contribution in school performance across the intervention period and grades. As a result, intervention effects were calculated by means of multilevel modelling (Shadish, Kyse, & Rindskopf, 2013; Van Den Noortgate & Onghena, 2003).

Following the modeling approach of Van Geel et. al. (2016), growth was modeled by modeling heterogeneity in (average) student achievement, while accounting for differences between measurement occasions in the different grade years in average test performance over students and schools. The differences in average achievements over grades were modeled as fixed effects, and student achievement and school achievement were allowed to vary across the general mean by introducing student and school-specific random effects.

Let *t* refers to the measurement occasion, *g* to the defined grade groups, *i* the student, and *j* the school. Then, random effects, represented as δ_{gij} $(g=1,2,3)$ were introduced for average achievement at baseline$(\text{class } g=1)$, over grades one to three $(\text{class } g=2)$ and grades four to six $(\text{class } g=3)$ at the student level. Then, the level-1 part of the model is represented by

$$Y_{tgij} = \mu_{tg} + \delta_{gij} + e_{tgij}, \quad (1)$$

where μ_{tg} represents the average score on occasion *t* in grade class *g*, and e_{tgij} is normally distributed with mean zero and residual variance σ^2. At the school level, the random effect β_{1j}, represents the effect of the intervention of schools and the β_{0j} represents the baseline performance of schools. The level-2 part, for each class *g*, is given by

$$\begin{aligned} \delta_{1ij} &= \beta_{0j} + \beta_{1j} Int_{1ij} + u_{1ij} \\ \delta_{2ij} &= \beta_{0j} + \beta_{1j} Int_{2ij} + u_{2ij} \\ \delta_{3ij} &= \beta_{0j} + \beta_{1j} Int_{3ij} + u_{3ij} \end{aligned} \quad (2)$$

where Int_{gij} represents the intervention variable and equals one when student *i* in school *j* and class grade *g* is measured during the intervention period (i.e., school *j* is participating in the intervention), and otherwise equals zero. The random effect δ_{gij} is only measured when student *i* has two or more measurements in class *g*, otherwise there is no random effect calculated for this student. This shows that the level-1 random effect representation provides sufficient flexibility to model the growth in student performance, while many students are only measured at some grades. The error component $\mathbf{u}_{ij}$ is assumed to be multivariate normally distributed and represents the random deviations of the individual latent growth measurement from the school-average performance for all class grades. Finally, the level-3 part of the model, represents the variation in school performances before and during the intervention across grade classes. The school-level random effects are assumed to be multivariate normally distributed,

$$\begin{aligned}\beta_{0j} &= \gamma_{00} + r_{0j} \\ \beta_{1j} &= \gamma_{10} + r_{1j},\end{aligned} \tag{3}$$

where γ_{00} is restricted to be zero, when including all occasion-specific effects μ_{tg} in the model. The error term r_{0j} represents the variation in school performances before the intervention, given the population-average occasion-specific performances μ_{tg}. The γ_{10} represents the population average intervention effect and r_{1j} represents the random deviation of school *j* of the population-average intervention effect. This error term is assumed to be normally distributed with mean zero, and the variance represents the variation in intervention effects across schools.

The analyses for measuring changes in mathematics and spelling performance were performed using the *lme4* package (Bates, Mächler, Bolker, & Walker, 2015) in R (RCoreTeam, 2013). Restricted maximum likelihood estimates (REML) were computed to estimate the model parameters. As mathematics and spelling were measured on different scales, the analyses for these two subjects were performed separately.

Interpretation of effects. The average difference between student scores at two subsequent test moments on the vertical latent scale was approximately 7.7 for mathematics (Cito, 2009a) and 3.3 for spelling (Cito, 2009b). Based on the fact that there were approximately five months of schooling between two test occasions, an effect of 1.54 (7.7/5; mathematics) and 0.66 (3.3/5; spelling) can be interpreted as the average increase in performance due to one additional month of schooling. This effect was expected to differ slightly between lower and upper grades, since the estimated differences in ability scores between two test occasions were larger in lower grades.

Results

In Figure 3, boxplots of the ability scores for mathematics achievement per grade are presented for the two years prior to the intervention and the two intervention years. As would be expected, ability scores improved over a student's school career. As displayed in Figure 3, the mean ability scores prior to the intervention tended to be slightly lower, compared to mean ability scores during the intervention. Boxplots of the ability scores for spelling (*see* Figure 4) revealed the same trend. Note that the ability scores for spelling and mathematics were not measured on the same scale and scores are thus incomparable.

[INSERT FIGURE 3 & FIGURE 4 ABOUT HERE]

Linear mixed effect analysis provides more insight into whether the differences in mean scores indicate an intervention effect. In the following section, first the results for mathematics are given, followed by the results for spelling.

Linear Mixed Effects Analysis for Mathematics

A total of seven models was analyzed, the baseline model included dummy variables representing the average performances per test occasion throughout a students' school career.

In the following, student background characteristics (Model 1), school characteristics (Model 2), a fixed intervention effect (Model 3), a random intervention effect (Model 4) and intervention trajectory (Model 5) were added. Finally Model 6 and Model 7 included interaction effects of intervention with a) trajectory (Model 6) and b) school and student SES with the intervention (Model 7). In Table 4, the results of the four most explanatory models are presented. As assessed through the decrease in information criteria values (i.e., AIC, BIC, Deviance), each subsequent model was a significant ($p < .05$) improvement over the previous one. The exception was Model 5 compared to Model 4 where 'intervention trajectory' was included as a fixed effect in the model. However, inclusion of an interaction effect of the intervention with trajectory revealed a significant improvement of the model ($\Delta\chi^2$ = 15.90, 2 df, $p < .001$).

The fixed effects in the baseline model showed that, in line with the boxplots in Figure 3, on average students improved their performances across assessments. The random intercept effect at the student level (i.e. halfway through grade three) and the random effects of grades 1-3 (grade class 2) and grades 4-6 (grade class 3) showed considerable variation between students' mathematics achievement scores at the first assessment in grade three. The student random effects of grade class 2 and 3 were strongly correlated (r = .85). This shows that the random effects are a consistent measure of student performance. The random effect of grade class 2 explained 10% of the variance in the student scores, of grade class 3 22% of the variance. The (conditional) intraclass correlation (ICC), conditional on the variance explained by the student random effects and average grade differences, was around 23% (i.e., dividing the school variance (13.28) by the sum of the school and the residual variance). This ICC of 23% indicate that conditional on student growth and grade-average differences, 23% of the variance in the student scores was explained by school differences.

The influence of student characteristics and school characteristics on mathematic achievement. Prior to testing the hypotheses, student characteristics were included in Model 1 and school characteristics were included in Model 2. Results indicated that high-SES students achieved higher mathematics outcomes compared to low-SES and medium-SES students. Moreover, boys tend to reach higher mathematics outcomes compared to girls. Regarding age, a positive effect of .53 (*Sd* = .02) was found, suggesting that the older a student was compared to his/her classmates, the higher his/her mathematics achievements.

At the school level, student achievement was lower in small schools compared to medium and large schools. Moreover, schools in urban and suburban areas performed poorer compared to schools in rural areas. Finally, the more low-SES students in a school, the lower the average mathematics achievement of these schools was.

The effect of the intervention on mathematics outcomes. To test the first hypothesis, concerning the effects of the intervention on student outcomes, a fixed effect of the intervention was included in Model 3. Results indicated that the general average intervention effect equaled 1.17, and differed significantly from zero, providing support for hypothesis 1. Subsequently, in Model 4 the random effects of the intervention was included to test whether intervention effects differed between schools (hypothesis 3). This resulted in an increase of the fixed intervention effect to 1.36. Moreover, the random effect of 3.41 revealed that the intervention effect indeed differed between schools. A likelihood ratio test on the random intervention effect revealed a significant result (p< .001), which showed that the intervention effect for mathematics varied across schools. Based on the 95% confidence interval of intervention effects in the population which ranges from -2.26 to 4.98 (1.36 ± 1.96*√3.41), we can thus conclude that the effect of the intervention is not positive for all participating schools. This is graphically illustrated in Figure 5, where the random intervention effect is plotted against the random intercept. In schools placed on the left side of the 0-axis, no effects of the intervention were observed, whereas in schools on the right side of the 0-axis, intervention effects were positive. Note that schools on the left side of the 0-axis prior to the intervention, generally achieved higher outcomes on mathematics compared to schools with large intervention effects.

[INSERT FIGURE 5 ABOUT HERE]

The influence of trajectory. For illustrative purposes, in Figure 5 schools are marked based on their trajectory. In line with our expectation (hypothesis 3), it seems that schools which focused on mathematics for more than one year showed the greatest improvement in achievement compared to schools which included mathematics only in year one or in year two. To investigate whether trajectory indeed explained part of the differences between schools (hypothesis 3), in Model 5 'trajectory' was included and in Model 6 an interaction effect of trajectory with the intervention effect was included. Although the inclusion of a main effect of trajectory did not lead to an improved model fit, according to the information criteria values ($\Delta\chi^2$ = 4.09, 2 df, p = .13), findings showed a significant effect of trajectory, suggesting that schools which worked on mathematics for more than one year, initially scored lower than schools which focused on mathematics in only the first year. By including the interaction effect of trajectory with the intervention in Model 6, we found that the intervention effect was largest for these schools. In sum, the intervention effect was largest for schools working on DBDM for math for two years, which were the schools who initially scored lower on math. Note that, due to the inclusion of the interaction effect with trajectory, the main effect of the intervention was no longer significant in Model 6.

Interactions with student-SES and school-SES. The final two hypotheses (hypothesis 4 and 5) were about whether the school-SES and student-SES could serve as an explanation for the differences in intervention effects. A higher intervention effect was expected in schools with a high proportion of low-SES students, and for students with low-SES or medium-SES backgrounds. Thus, the intervention effect was expected to differ between schools and students with different SES scores. In Model 7, these interaction-effects were included. At the school level, a significant interaction effect was found for the estimated adjustment of the intervention effect for all students of schools with on average low-SES students. The significant interaction effect at the school level suggests that the intervention effect was larger in schools with a high proportion of low-SES students compared to schools with a high proportion of medium-SES and high-SES students. This supported hypothesis 4. Conditional on this school-level interaction effect, the intervention effect was significantly higher for students with low-SES or high-SES backgrounds compared to students with medium-SES backgrounds. At the individual level, both low-SES students as well as high-SES students benefitted more from the intervention compared to medium-SES students. This is not in line with our expectations, as it was not expected that high-SES students would also benefit more from the intervention than medium-SES students did. Thus, hypothesis 5 was only partially supported.

Linear Mixed Effects Analysis for Spelling

Results of the linear mixed effects analysis for spelling achievement are provided in Table 5. A total of 28 schools were included in the analyses of spelling achievement. Similar to the mathematics analysis, seven models were used to analyze the data. Not all subsequent models led to significant improvements.

In the baseline model, the fixed effects of the subsequent test occasions showed a similar growth pattern as the pattern shown in Figure 4. The variance components revealed much variation at the student level, whereas the clustering of grades 1-3 explained 15% of the total variance and the clustering of grades 4-6 explained 31% of the total variance. The (conditional) intra-class correlation, conditional on the student random effects and average grade differences was around 13%, representing the percentage of (conditional) variance in student scores explained by the schools. The student random effects at grade class 2 and 3 correlated around .96, and shows that a general measure of performance underlies these random effect measurements.

The influence of student characteristics and school characteristics on spelling achievement. In Model 1, student characteristics were added. Findings showed that high-SES students performed higher on spelling compared to medium-SES and low-SES students. Moreover, girls tended to achieve higher spelling scores compared to boys. Finally, similarly to mathematics achievement, older students performed better on spelling compared to their younger peers. Findings of Model 2, in which school characteristics were added, showed that spelling achievement in small schools was on average lower compared to medium sized schools. No effects of urbanization or school-SES on spelling achievements were found. Therefore, in the subsequent models, urbanization was excluded from the model. The main effect of school-SES remained in the model in order to test hypothesis 4 in a later model.

The effect of the intervention on spelling outcomes In order to test hypothesis 1, a fixed intervention effect was included in Model 3. Findings showed a significant effect of .71, suggesting that, on average, the intervention had a positive effect on spelling outcomes. To test whether this effect differed between schools, a random intervention effect was specified in Model 4, resulting in a random intervention effect of .85. The likelihood ratio test on the random intervention effect also revealed a significant result (p< .001), which showed that the intervention effect varied across schools. By adding the random intervention effect, the fixed intervention effect slightly increased to .79. The 95% confidence interval of intervention effects in the population ranged from -1.02 to 2.60 (.79 ± 1.96*√.85), revealing that not all participating schools experienced positive effects. In Figure 6, this is illustrated. As can be seen in the graph, in the majority of schools the random intervention effects were positive, but 7 out of the 28 schools did not achieve higher student achievement growth during the two intervention years.

[INSERT FIGURE 6 ABOUT HERE]

The influence of trajectory. As can be seen in Figure 6, only three schools focused on spelling during more than one intervention year. For this reason, results from Model 5 and 6 should be interpreted with caution. In Model 5, the fixed trajectory effect was included, this effect was not significant. Furthermore, the model fit did not improve due to inclusion of trajectory. To test hypothesis 3, an interaction effect with intervention was included in Model 6. This resulted in a positive interaction effect for schools which focused on spelling for more than one year, however, note that this was based on three schools. For the three schools focusing on spelling for more than one year, the intervention effect was substantially higher than for schools which focused on spelling for only one year.

The influence of socioeconomic status. Finally, to test hypotheses 4 and 5 regarding school-SES and student-SES, interaction effects between SES and the intervention effect were included in Model 7. None of these interaction effects were significant, nor led to an improvement of the model fit. Therefore, hypotheses 4 and 5 could not be supported for spelling outcomes.

Evaluating Model Fit

In order to test model assumptions, several analyses were conducted. First, fitted scores and residuals at the student level (level 1) were plotted to check for outliers, and observed scores were plotted against fitted scores to check for systematic patterns of misfit. No abnormal patterns or outliers were found in the spelling data.

Next, for each random effect at the first level, the ordered fitted random effect residuals were plotted against their normal quantiles in a normal QQ-plot. For both, mathematics and spelling, the random intercept effects followed an approximately normal distribution, but student level random effects, namely grades 1-3 and grades 4-6, showed more deviations. Both random effects seemed to be more peaked than what would be expected under the normal distribution. This means that we measured less variation in student scores in grades 1-3 and grades 4-6 than expected under the normal distribution. This can partly be explained by the fact that almost 30% of the students were measured only three or fewer times, which made it difficult to distinguish their average performance. As a result, average student performance showed less variation than expected under the normal distribution, leading to more peaked random effect distributions. Furthermore, the sample showed fewer students with more extreme average grade scores than expected under normality, but the average grade-score distribution could still be normal in the population.

At the school level (level 2), QQ-plots were examined for each random effect (random intercept and random intervention effects). For mathematics, both the random intercept residuals and the school-level random intervention effects showed some deviations. A similar pattern was found for spelling. However, there were too few schools in the sample to make inferences about the normality assumption at this level.

Finally, to test the assumption of homoscedasticity of Level-1 residual variances, the Chi-square test developed by Snijders and Bosker (1999, pp. 126–127) was used. The Chi-square test revealed significant results for both mathematics and spelling, indicating that at least one of the school-specific level-1 residual variances was significantly different from the other schools. Therefore, the logarithm of the estimated variances were plotted to evaluate the variation across schools. The 95% lower and upper-bound were computed, assuming that the logarithm of residual variances were normally distributed. It became apparent that a few outliers led to a significant test result. Furthermore, the high number of students per school also led to significant differences, since many of the school-specific level-1 variances were estimated very accurately. Although this provided support for the modelling of school-specific residual variances, for most of the schools the assumption of a common residual variance was acceptable.

[INSERT TABLE 4 & TABLE 5 ABOUT HERE]

Conclusions and Discussion

Although the concept of DBDM is applicable to all school subjects, studies into the effects of DBDM interventions on student achievement are merely aimed at mathematics or reading. In order to broaden our understanding of the link between DBDM interventions and student performance, the effect of a DBDM intervention on both mathematics as well as spelling was studied.

We argued that to be able to influence student performance regardless of the subject-matter content, two characteristics of a DBDM intervention are specifically important. First, DBDM interventions should include all DBDM components in a coherent and consistent way. Second, interventions should take both the school level as well as the teacher level into account. In line with these criteria, we developed a two-year DBDM intervention. In a previous study, this intervention proved to be effective for mathematics. To broaden our understanding of the effectiveness of this intervention and to increase the generalizability of the findings, the effects of the intervention on student achievement were analyzed in a new cohort of 40 schools. In this study, we were not only interested in the intervention effects on mathematics achievement but also in the effects on spelling achievement.

In Table 6 the results of testing the hypotheses for both mathematics and spelling are summarized.

[INSERT TABLE 6 ABOUT HERE]

It was expected that student achievement growth for both mathematics as well as for spelling would be greater during the intervention, compared to student achievement growth in the two years prior to the intervention. Our findings showed that this was indeed the case for both mathematics as well as for spelling (Hypothesis 1, supported). On average an effect of respectively 1.39 for mathematics and .79 for spelling was found. Considering that an effect of 1.54 (mathematics) and .66 (spelling) on average can be interpreted as the expected increase in performance due to one additional month of schooling, it seems that for both subjects the intervention resulted in approximately one extra month of schooling. However, findings also revealed that the intervention-effects differed between schools (Hypothesis 2, supported). There were schools for which the intervention effects were much larger than average. In contrast, for spelling in 25% and for mathematics in 30% of the schools, the intervention did not have an effect on student achievement growth.

We hypothesized that the trajectory schools chose to follow during the intervention would influence the effects of the intervention. Prior to the intervention, schools could choose with which subject they wanted to start the intervention, and in the second year schools could decide to focus on another subject, or stick to the same subject. In line with the literature suggesting that it takes two or more years to accomplish achievement effects resulting from a professional development trajectory (Desimone, 2009; Timperley, 2008), we expected the largest results for those schools that worked on the same subject during the entire two-year intervention. For mathematics, findings confirmed this hypothesis (hypothesis 3, supported). Schools which started with mathematics and remained focused on mathematics in the second year, showed greater improvement in student performance than schools which changed their focus. Based on Figure 5 it could be argued that these schools were also the schools with the lowest mathematics starting scores. For spelling, only three schools remained focused on spelling for more than one year. Although these three schools also showed higher intervention effects on average, compared to the other schools, the small size of this group of schools offers too little evidence for a claim that this holds for the spelling trajectory in general.

Furthermore, it was expected that school-SES would explain part of the differences in intervention effects across schools. For mathematics, the intervention effects were largest for schools with a large proportion of low-SES students. For spelling, the intervention effects did not significantly differ between low-SES, medium-SES or high-SES schools. Thus, hypothesis 4 is supported for mathematics, but not for spelling. Additionally, at the student level we expected the same trend as for hypothesis 4: students from a low-SES background would benefit most from the intervention. The results of the spelling analysis did not confirm this hypothesis. However, for mathematics we found that student outcomes improved to a greater extent for low-SES and high-SES students in comparison to medium-SES students. Therefore, hypothesis 5 is supported for mathematics but not for spelling.

This study underlines the finding that mathematics outcomes improved especially for low-SES schools of the first study. The effect sizes for mathematics are comparable to the effect sizes found in the previous study. In contrast to the first study, a significant effect of intervention trajectory was found, in favor of those schools who focused on mathematics for more than one year. It is possible that this was caused by the fact that in the current study, schools could choose their subject in the first intervention year whereas in the first study, schools were required to start with mathematics. It is likely that schools that chose mathematics performed poor on this subject, whereas in the first study schools may have performed quite well on mathematics and, as a result, showed less improvement in performance.

Moreover, this study built upon the findings of the previous study by showing that the intervention can have a positive effect on spelling. It looks like the DBDM is not subject-specific and can be implemented in multiple subjects. Interestingly, the effect of the intervention was not related to students' SES for spelling. We expected the intervention to be especially beneficial for students from low socio-economic backgrounds as these students have a higher risk of becoming poor spellers. Nevertheless, the DBDM-intervention seemed to improve students' spelling outcomes regardless of their socio-economic background.

The self-selection of schools complicates a generalization of the results to a larger population. However, the multilevel modeling approach combines the intervention results of all participating schools. This makes it possible to make joint inferences for all schools in the sample, and they cover a substantial part of the population of schools (e.g., compared to all Dutch schools, participating schools had relatively more students from a lower-SES background). More generally, the demographic information of the schools in the sample showed that they did not misrepresent to a larger extent specific subpopulations. The multilevel approach avoids the typical issue of single-subjects studies by providing a way to combine the results of the intervention studies at the participating schools. The estimated average intervention effect is a result of combining the intervention studies.

In the single-subject design, participating schools served as their own control in the pre-intervention period. This made it possible to estimate school-specific intervention effects, and the intra-school measurements showed that valid results were obtained by checking the criteria of (internal) validity. By pooling information across studies more reliable and accurate information was obtained about the effects of the intervention study.

Future Research

In this study, the data did not allow us to model student performance per grade level. This would have led to a huge number of random effect parameters and a complex missing data problem, since many students were not measured more than once at each grade level. This problem was avoided by measuring average student performance in grades 1-3 and 4-6. As a result, we could not study differences in intervention effects across grades. In future work, it would be interesting to disentangle the average school intervention effect into grade-specific effects, to deepen our understanding of DBDM.

Furthermore, it would be interesting to conduct a retention measure in the intervention schools. As Desimone (2009) argued, it takes time to implement reform within a school and that fully changing practices in schools can take up to five years. During the two intervention years, student achievement growth improved with approximately one month of extra schooling. The question remains as to how these effects develop after these two years. A follow-up study can provide insight into the development of the effect over time: will it remain stable, increase, or decrease?

Another question raised is how we can explain the differences in effects found between spelling and mathematics. Whereas school-SES and intervention trajectory proved to explain differences in intervention effects between schools for mathematics, such relationships were not found for spelling. More research into DBDM and spelling (and especially the relationship between school-SES, trajectory and spelling achievement) is needed to clarify these findings.

Additionally, although school-SES and intervention trajectory seem to explain some of the variation in intervention effects for mathematics, a more detailed analysis of the *implementation process* in the participating schools could provide further insight into the reasons for these differences. The DBDM literature includes a broad range of factors that potentially explain differences in DBDM effects (Ikemoto & Marsh, 2007; Mandinach, 2012; Schildkamp, Karbautzki, & Vanhoof, 2014; Schildkamp & Lai, 2013a; Schildkamp & Poortman, 2015; Visscher & Ehren, 2011), including teachers' attitudes, leadership, internal collaboration and the quality of the teachers. In future research the explanatory power of these factors has to be investigated in depth.

Finally, in this study we claimed the importance of taking into account the teacher level in order to improve student outcomes. To support teachers with the final step of the DBDM-process - execute planned strategies in their classrooms to reach the goals set based on their data analysis - two coaching sessions per school were planned in the second intervention year. Although these sessions were valued by both participating teachers as well as trainers, it cannot be assumed these coaching sessions were decisive for the effectiveness of the intervention as the coaching was not the key feature of the intervention and literature suggest that coaching processes should be more intensive (individual interaction between coach and teacher at least every couple weeks) and sustained (teachers receive coaching over a longer period of time) (Kraft, Blazar, & Hogan, 2018). In future research into DBDM-interventions we therefore recommend to support teachers in their classroom more frequent and in a more systematic way.

References

Bates, D., Mächler, M., Bolker, B., & Walker, S. (2015). Fitting Linear Mixed-Effects Models Using lme4. *Journal of Statistical Software*, *67*(1). http://doi.org/10.18637/jss.v067.i01

Birman, B. F., Desimone, L. M., Porter, A. C., & Garet, M. S. (2000). Designing professional development that works. *Educational Leadership*, *57*(8), 28–33.

Black, B. P., & Wiliam, D. (1998). Inside the Black Box: Raising Standards Through Classroom Assessment. *Assessment in Education*, *5*(1), 7–74.

Boudett, K. P., City, E., & Murnane, R. (2005). *Data Wise: A step-by-step guide to using assessment results to improve teaching and learning*. Cambridge, MA: Harvard Education Press.

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. http://doi.org/10.3102/0162373711412765

Cito. (2009a). *Rekenen-Wiskunde Handleiding*. Arnhem.

Cito. (2009b). *Spelling Handleiding*. Arnhem.

Coburn, C. E., & Turner, E. O. (2012). The Practice of Data Use : An Introduction. *American Journal of Education*, *118*(2), 99–111. http://doi.org/10.1086/663272

Darling-Hammond, L. (2000). Teacher quality and student achievement: A review of state policy evidence. *Education Policy Analysis Archives*. http://doi.org/10.1038/sj.clp

De Wijs, A., Kamphuis, F., Kleintjes, F., & Tomessen, M. (2010). *Wetenschappelijke verantwoording: spelling voor groep 3 tot en met 6 [scientific justification: spelling for grade 3 to 6]*. Arnhem: Cito. Retrieved from https://docplayer.nl/17762101-Wetenschappelijke-verantwoording-spelling-voor-groep-3-tot-en-met-6.html

Desimone, L. M. (2009). Improving impact studies of teachers' professional development: Toward better conceptualizations and measures. *Educational Researcher, 38*(3), 181–199.

Flay, B. R., & Collins, L. M. (2005). Historical Review of School-Based Randomized Trials for Evaluating Problem Behavior Prevention Programs. *The ANNALS of the American Academy of Political and Social Science, 599*(1), 115–146. http://doi.org/10.1177/0002716205274941

Fuchs, L. S., & Fuchs, D. (1986). Effects of systematic formative evaluation: A meta-analysis. *Exceptional Children*, *53*(199–208).

Garet, M. S., Porter, A. C., Desimone, L. M., Birman, B. F., & Yoon, K. S. (2001). What makes professional development effective? Results from a national sample of teachers. *American Educational Research Journal*, *38*(4), 915–945. Retrieved from http://www.jstor.org/stable/3202507

Graham, S., Morphy, P., Harris, K. R., Fink-Chorzempa, B., Saddler, B., Moran, S., & Mason, L. (2008). Teaching Spelling in the Primary Grades: A National Survey of Instructional Practices and Adaptations. *American Educational Research Journal, 45*(3), 796–825. http://doi.org/10.3102/0002831208319722

Graham, S., & Santangelo, T. (2014). Does spelling instruction make students better spellers, readers, and writers? A meta-analytic review. *Reading and Writing, 27*(9), 1703–1743. http://doi.org/10.1007/s11145-014-9517-0

Handley, M. A., Lyles, C. R., McCulloch, C., & Cattamanchi, A. (2018). Selecting and Improving Quasi-Experimental Designs in Effectiveness and Implementation Research. *Annual Review of Public Health*, *39*(1), 5–25. http://doi.org/10.1146/annurev-publhealth-040617-014128

Hattie, J. (2009). *Visible Learning: A Synthesis of over 800 Meta-Analyses relating to*

Achievement. London: Routledge.

Hattie, J., & Timperley, H. (2007). The Power of Feedback. *Review of Educational Research, 77*(1), 81–112. http://doi.org/10.3102/003465430298487

Henderson, S., Petrosiono, A., Guckenburg, S., & Hamilton, S. (2007). *Measuring how benchmark assessments af fect student achievement Measuring how benchmark assessments affect student achievement*. (Issues & Answers Report, REL 2007-No. 039). Washington, DC. Retrieved from http://ies.ed.gov/ncee/edlabs

Ikemoto, G. S., & Marsh, J. A. (2007). Cutting Through the “data-driven” mantra: Different conceptions of data-driven decision making. In *Yearbook of the National Society for the Study of Education* (Vol. 106, pp. 105–131). http://doi.org/10.1111/j.1744-7984.2007.00099.x

Inspectie van het Onderwijs. (2012). *Beoordeling van opbrengsten in het basisonderwijs*. Utrecht, The Netherlands.

Inspectie van het Onderwijs. (2013). *Analyse en waardering van opbrengsten PO [Analysis and valuation of proceeds primary education]*. Utrecht.

Janssen, J., & Hickendorff, M. (2010). *Categorieënanalyse bij de lovs-toetsen rekenen-wiskunde [Categorical analysis for student monitoring math tests]*. Retrieved from https://docplayer.nl/20026129-Categorieenanalyse-bij-de-lovstoetsen.html

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 justification of the student monitoring math tests for grade 3 to 8]*. Arnhem: Cito.

Jenson, W. R., Clark, E., Kircher, J. C., & Kristjansson, S. D. (2007). Statistical reform: Evidence-based practice, meta-analyses, and single subject designs. *Psychology in the Schools, 44*(5), 483–493. http://doi.org/10.1002/pits.20240

Ji, P., DuBois, D. L., Flay, B. R., & Brechling, V. (2008). “Congratulations, you have been randomized into the control group!(?)”: Issues to consider when recruiting schools for matched-pair randomized control trials of prevention programs. *Journal of School Health*, *78*(3), 131–139. http://doi.org/10.1111/j.1746-1561.2007.00275.x

Kamphuis, F., & Moelands, F. (2000). A student monitoring system. *Educational Measurement: Issues and Practice*, 28–30.

Kaufman, T., Graham, C., Picciano, A., Popham, J. A., & Wiley, D. (2014). Data-Driven Decision Making in the K-12 Classroom. In J. M. Spector, M. D. Merrill, J. Elen, & M. J. Bishop (Eds.), *Handbook of Research on Educational Communications and Technology SE - 27* (pp. 337–346). Springer New York. http://doi.org/10.1007/978-1-4614-3185-5_27

Keuning, T., Van Geel, M., Fox, J.-P., & Visscher, A. J. (2016). The Transformation of Schools’ Social Networks During a Data-Based Decision Making Reform. *Teachers College Record*, *118*(090308).

Kingston, N., & Nash, B. (2011). Formative assessment: A meta-analysis and a call for research. *Educational Measurement: Issues and Practice*, *30*(4), 28–37. http://doi.org/10.1111/j.1745-3992.2011.00220.x

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*, *199*(2), 254–284.

Konstantopoulos, S., Miller, S. R., & van der Ploeg, a. (2013). The Impact of Indiana’s System of Interim Assessments on Mathematics and Reading Achievement. *Educational Evaluation and Policy Analysis*, *35*(4), 481–499. http://doi.org/10.3102/0162373713498930

Kraft, M. A., Blazar, D., & Hogan, D. (2018). The Effect of Teacher Coaching on Instruction and Achievement: A Meta-Analysis of the Causal Evidence. *Review of Educational Research*, *XX*(X), 003465431875926. http://doi.org/10.3102/0034654318759268

Kratochwill, T. R., Hitchcock, J., Horner, R. H., Levin, J. R., Odom, S. L., Rindskopf, D. M., & Shadish, W. R. (2010). *Single-case designs technical documentation*. Retrieved from http://ies.ed.gov/ncee/wwc/pdf/wwc_scd.pdf

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: Springer Netherlands. http://doi.org/10.1007/978-94-007-4816-3

Love, N., Stiles, K. E., Mundry, S., & DiRanna, K. (2008). *A data coach's guide to improve learning for all students: Unleashing the power of collaborative inquiry*. Thousand Oaks, CA: Corwin Press.

Lumpe, A. T. (2007). Research-based professional development: Teachers engaged in professional learning communities. *Journal of Science Teacher Education*, *18*(1), 125–128. http://doi.org/10.1007/s10972-006-9018-3

Makel, M. C., & Plucker, J. a. (2014). Facts Are More Important Than Novelty: Replication in the Education Sciences. *Educational Researcher*, *43*(6), 304–316. http://doi.org/10.3102/0013189X14545513

Mandinach, E. B. (2012). A perfect time for data use: Using data-driven decision making to inform practice. *Educational Psychologist*, *47*(2), 71–85. http://doi.org/10.1080/00461520.2012.667064

Mandinach, E. B., & Gummer, E. S. (2015). Data-driven decision making: components of the enculturation of data use in education. *Teachers College Record*, *117*(4).

Marsh, J. A., Pane, J. F., & Hamilton, L. S. (2006). *Making sense of data-driven decision making in education*. Santa Monica, CA.

McCaffrey, D. F., & Hamilton, L. S. (2007). *Value-added assessment in practise: Lessons from the Pennsylvania value-added assessment system pilot project*. Santa Monica, CA: RAND Corporation. Retrieved from http://www.rand.org/pubs/technical_reports/TR506

Mullis, I. V. S., Martin, M. O., Foy, P., & Arora, A. (2012). *TIMSS 2011 international results in mathematics*. Chestnut Hill, MA.

Nye, B., Konstantopoulos, S., & Hedges, L. V. (2004). How Large Are Teacher Effects ? *Educational Evaluation and Policy Analysis*, *26*(3), 237–257.

Oláh, L. N., Lawrence, N. R., & Riggan, M. (2010, April 26). Learning to Learn From Benchmark Assessment Data: How Teachers Analyze Results. *Peabody Journal of Education*. http://doi.org/10.1080/01619561003688688

Orland, M. (2015). Research and Policy Perspectives on Data - Based Decision Making in Education. *Teachers College Record*, *117*(4).

Quint, J. C., Sepanik, S., & Smith, J. K. (2008). *Using Student Data to Improve Teaching and Learning: Findings from an Evaluation of the Formative Assessments of Student Thinking in Reading (FAST-R) Program in Boston Elementary Schools*. New York, NY: MDRC.

RCoreTeam. (2013). R: A language and environment for statistical computing. Vienna, Austria: R Foundation for Statistical Computing. Retrieved from http://www.r-project.org/

Ritzema, E. S. (2015). *Professional development in data use: The effects of primary school teacher training on teaching practices and students' mathematical proficiency*. University of Groningen.

Schildkamp, K., Ehren, M., & Lai, M. K. (2012). Editorial article for the special issue on data-based decision making around the world: from policy to practice to results. *School Effectiveness and School Improvement*, *23*(2), 123–131.

http://doi.org/10.1080/09243453.2011.652122

Schildkamp, K., Karbautzki, L., & Vanhoof, J. (2014). Exploring data use practices around Europe: Identifying enablers and barriers. *Studies in Educational Evaluation*, *42*, 15–24. http://doi.org/10.1016/j.stueduc.2013.10.007

Schildkamp, K., & Lai, M. K. (2013a). Conclusion and a data use framework. In K. Schildkamp, M. K. Lai, & L. Earl (Eds.), *Data-based Decision Making in Education: challenges and opportunities* (pp. 177–191). Dordrecht: Springer. http://doi.org/10.1007/978-94-007-4816-3

Schildkamp, K., & Lai, M. K. (2013b). Introduction. In K. Schildkamp, M. K. Lai, & L. Earl (Eds.), *Data-based Decision Making in Education: challenges and opportunities* (pp. 1–7). Dordrecht: Springer. http://doi.org/10.1007/978-94-007-4816-3

Schildkamp, K., & Poortman, C. L. (2015). Factors influencing the function of data teams. *Teachers College Record*, *117*(4). Retrieved from http://www.tcrecord.org/content.asp?contentid=17851

Schildkamp, K., Poortman, C. L., & Handelzalts, A. (2015). Data teams for school improvement. *School Effectiveness and School Improvement: An International Journal of Research, Policy and Practice*. http://doi.org/10.1080/09243453.2015.1056192

Schmidt, S. (2009). Shall we really do it again? The powerful concept of replication is neglected in the social sciences. *Review of General Psychology*, *13*(2), 90–100. http://doi.org/10.1037/a0015108

Shadish, W. R., Cook, T. D., & Campbell, D. T. (2002). *Experimental and quasi-experimental designs for generalized causal inference*. Bostan, MA: Houghton Mifflin.

Shadish, W. R., Kyse, E. N., & Rindskopf, D. M. (2013). Analyzing data from single-case designs using multilevel models: New applications and some agenda items for future research. *Psychological Methods*, *18*(3), 385–405. http://doi.org/10.1037/a0032964

Slavin, R. E., Cheung, A., Holmes, G., Madden, N. A., & Chamberlain, A. (2012). Effects of a Data-Driven District Reform Model on State Assessment Outcomes. *American Educational Research Journal*, *50*(2), 371–396. http://doi.org/10.3102/0002831212466909

Snijders, T. A. B., & Bosker, R. J. (1999). *Multilevel analysis: an introduction to basic and advanced multilevel modeling*. London, UK: Sage Publishers.

Supovitz, J. A. (2012). Getting at student understanding - The key to teachers' use of test data. *Teachers College Record*, *114*.

Timperley, H. (2008). *Teacher professional learning and development*. International Academy of Education.

Turner, E. O., & Coburn, C. E. (2012). Interventions to promote data use: An introduction. *Teachers College Record*, *114*(110301), 13.

Van Den Noortgate, W., & Onghena, P. (2003). Multilevel meta-analysis: A comparison with traditional meta-analytical procedures. *Educational and Psychological Measurement*, *63*, 765–790. http://doi.org/10.1177/0013164403251027

Van der Kleij, F. M., Vermeulen, J. A., Schildkamp, K., & Eggen, T. J. H. M. (2015). Integrating data-based decision making, Assessment for Learning and diagnostic testing in formative assessment. *Assessment in Education: Principles, Policy & Practice*, (February 2015), 1–20. http://doi.org/10.1080/0969594X.2014.999024

van Geel, M., Keuning, T., Visscher, A. J., & Fox, J.-P. (2016). Assessing the Effects of a School-Wide Data-Based Decision-Making Intervention on Student Achievement Growth in Primary Schools. *American Educational Research Journal*, *53*(2), 360–394. http://doi.org/10.3102/0002831216637346

Van Veen, K., Zwart, R., & Meirink, J. (2011). What makes teacher professional development effective? A literature review. In M. Kooy & K. Van Veen (Eds.), *Teacher learning that* *matters* (pp. 3–21). New York: Routledge.

Visscher, A. J., & Ehren, M. (2011). De eenvoud en complexiteit van Opbrengstgericht Werken. [The simplicity and complexity of data-driven teaching], 1–37. Retrieved from http://www.rijksoverheid.nl/documenten-en-publicaties/rapporten/2011/07/13/de-eenvoud-en-complexiteit-van-opbrengstgericht-werken.html

Vlug, K. F. M. (1997). Because every pupil counts: the success of the pupil monitoring system in The Netherlands. *Education and Information Technologies*, *2*(4), 287–306. http://doi.org/10.1023/A:1018629701040

Table 1

Project overview

		Type of meeting	Content description
Y1		*School leader / school board meeting*	Fulfilling practical preconditions and stressing the importance of the role of the school leader/school board
	1.1	Team meeting	Analyzing test score data from the student monitoring system
	1.2	Team meeting	Subject matter content – curriculum Individual diagnosis of students' learning needs
	1.3	Team meeting	Goal setting and developing instructional plans
	1.4	Team meeting	Instructional plans in practice Monitoring and adjusting instructional plans based on test data from content mastery tests and daily work in class
		School leader / school board meeting	Discussing progress and the goals for the next period (trainer, school leader and school board)
	1.5	Team meeting	Team meeting: evaluating standardized test performance data
	1.6	Team meeting	Collaboration in the school: how to learn from each other by using classroom observations
		School leader / school board meeting	Discussing progress and goals for the next period (trainer, school leader and school board)
	1.7	Team meeting	Team meeting: evaluating standardized test performance data
Y2	2.1	Team meeting	Option 1: Continue with the same subject: based on issues raised by school Option 2: New subject: tests and analysis for new subject, content and curriculum
	2.2	Classroom observations	Classroom observations
	2.3	Team meeting	Team meeting: evaluating standardized test performance data Optional: adding another subject, tests and analysis for new subject
	2.4	Classroom observations	Classroom observations
	2.5	Team meeting	Team meeting: evaluating standardized test performance data.

Table 2

School characteristics (N=39)

		N	(%)
School Size (number of students)	Small (<150)	13	(33.3%)
	Medium (150-350)	20	(51.3%)
	Large (>350)	6	(15.4%)
Urbanization	Rural	17	(43.6
	Suburban	15	(38.5%)
	Urban	7	(17.9%)
School SES	High	12	(30.8%)
	Medium	21	(53.8%)
	Low	6	(15.4%)
Main intervention subject	Mathematics	20	(51.3%)
	Spelling	15	(38.5%)
	Reading	3	(7.7%)
	Vocabulary	1	(2.6%)
Trajectory Spelling	No spelling at all	11	(38.2%)
	2nd year spelling	13	(33.3%)
	1st year spelling	12	(30.8%)
	> 1 year spelling	3	(7.7%)
Trajectory Mathematics	No mathematics at all	5	(12.8%)
	2nd year mathematics	14	(35.9%)
	1st year mathematics	11	(28.2%)
	> 1 year mathematics	9	(23.1%)

Table 3

Student characteristics for mathematics (n=8,023) and spelling (n=6,610)

		Mathematics		Spelling	
		N	(%)	N	(%)
Gender	Boy	4024	(50.2%)	3330	(50.4%)
	Girl	3999	(49.8%)	3280	(49.6%)
Student SES	High	6426	(80.1%)	5686	(86.0%)
	Medium	676	(8.4%)	443	(6.7%)
	Low	914	(11.4%)	474	(7.2%)
	Unknown	7	(0.1%)	7	(0.1%)
Number of observations per student	2	1754	(21.9%)	1215	(18.4%)
	3	622	(7.8%)	580	(8.8%)
	4	1464	(18.2%)	1097	(16.6%)
	5	590	(7.4%)	504	(7.6%)
	6	1141	(14.2%)	996	(15.1%)
	7	607	(7.6%)	592	(9.0%)
	8	1476	(18.4%)	1248	(18.9%)
	> 8	369	(4.6%)	378	(5.7%)

Table 4

Results of the linear mixed effects analysis for mathematic achievement

	Model 0		Model 4 Random intervention effect		Model 6 Trajectory* intervention		Model 7 SES* intervention	
	Est.	(SE)	Est.	(SE)	Est.	(SE)	Est.	(SE)
Intercept	28.12	(.66) ǂ	26.40	(1.05) ǂ	28.26	(1.12) ǂ	28.71	(1.13) ǂ
Student Level								
Test end grade 1	11.48	(.18) ǂ	11.89	(.18) ǂ	11.89	(.18) ǂ	11.89	(.18) ǂ
Test mid grade 2	20.21	(.19) ǂ	19.76	(.19) ǂ	19.76	(.19) ǂ	19.76	(.19) ǂ
Test end grade 2	31.32	(.19) ǂ	31.13	(.19) ǂ	31.13	(.19) ǂ	31.12	(.19) ǂ
Test mid grade 3	40.00	(.20) ǂ	39.08	(.21) ǂ	39.08	(.21) ǂ	39.08	(.21) ǂ
Test end grade 3	47.54	(.20) ǂ	46.87	(.21) ǂ	46.86	(.21) ǂ	46.86	(.21) ǂ
Test mid grade 4	54.42	(.22) ǂ	53.16	(.24) ǂ	53.15	(.24) ǂ	53.15	(.24) ǂ
Test end grade 4	60.47	(.22) ǂ	59.44	(.24) ǂ	59.44	(.24) ǂ	59.43	(.24) ǂ
Test mid grade 5	69.16	(.23) ǂ	67.50	(.26) ǂ	67.49	(.26) ǂ	67.49	(.26) ǂ
Test end grade 5	73.87	(.23) ǂ	72.38	(.26) ǂ	72.38	(.26) ǂ	72.37	(.26) ǂ
Test mid grade 6	81.62	(.26) ǂ	79.74	(.31) ǂ	79.74	(.31) ǂ	79.74	(.31) ǂ
Student SES – high			6.62	(.53) ǂ	6.62	(.53) ǂ	6.15	(.55) ǂ
Student SES – low			.25	(.67)	.25	(.67)	-.32	(.70)
Student gender (ref=boy)			-3.63	(.29) ǂ	-3.62	(.29) ǂ	-3.62	(.29) ǂ
Student age (months)			.51	(.02) ǂ	.51	(.02) ǂ	.51	(.02) ǂ
Intervention			1.36	(.34) ǂ	.51	(.46)	-.43	(.53)
Intervention * StudentSES high							.97	(.33) ǂ
Intervention * StudenSES low							1.20	(.41) ǂ
School Level								
SchoolSize – large			.63	(1.14)	.59	(1.08)	.61	(1.08)
SchoolSize – small			-1.99	(.94) ǂ	-2.32	(.91) ǂ	-2.31	(.91) ǂ
Suburban			-2.89	(.88) ǂ	-2.70	(.86) ǂ	-2.70	(.86) ǂ
Urban			-2.06	(1.12) *	-2.23	(1.12) *	-2.22	(1.11) *
SchoolSESlow			-2.42	(1.01) ǂ	-2.23	(.98) ǂ	-3.24	(1.09) ǂ
SchoolSEShigh			2.23	(1.02) ǂ	2.38	(1.03) ǂ	2.60	(1.12) ǂ
Trajectory: >1 year math (ref = 1^{st} year)					-4.29	(1.16) ǂ	-4.03	(1.15) ǂ
Trajectory: 2^{nd} year math (ref = 1^{st} year)					-1.75	(1.02) *	-1.63	(1.01)
Intervention * > 1 year math					2.84	(.68) ǂ	2.50	(.63) ǂ
Intervention *2^{nd} year math					.24	(.61)	.05	(.57)
Intervention * SchoolSESlow							1.32	(.62) ǂ
Intervention * SchoolSEShigh							-.28	(.59)
Variance Components								
Student level								

	Model 0		Model 4 Random intervention effect		Model 6 Trajectory* intervention		Model 7 SES* intervention	
	Est.	(SE)	Est.	(SE)	Est.	(SE)	Est.	(SE)
(Intercept)	168.91		172.55		172.61		172.63	
Clust123	34.73		36.98		37.06		37.01	
Clust456	74.26		71.23		71.47		71.19	
School level								
(Intercept)	13.28		7.32		4.81		4.57	
Intervention			3.41		1.96		1.55	
Residual	43.70		40.91		40.91		40.91	
Information Criteria								
AIC	295244.10		293770.40		293757.6		293749.9	
BIC	295407.80		294046.10		294067.7		294094.4	
-2LogLik	-147603.00		-146853.20		-146842.8		-146835	
Deviance	295206.10		293706.40		293685.6		293669.9	

*Note. * p < .05; ǂp < .01.*

Table 5

Results of the linear mixed effects analysis for spelling achievement

	Model 0		Model 4 Random intervention effect		Model 6 Trajectory* intervention		Model 7 SES* intervention	
	Est.	(SE)	Est.	(SE)	Est.	(SE)	Est.	(SE)
Intercept	107.50	(.30) ǂ	104.20	(.58) ǂ	104.20	(.81) ǂ	104.30	(.83) ǂ
Student Level								
Test end grade 1	6.26	(.11) ǂ	6.39	(.11) ǂ	6.39	(.11) ǂ	6.39	(.11) ǂ
Test mid grade 2	11.07	(.12) ǂ	10.95	(.12) ǂ	10.95	(.12) ǂ	10.95	(.12) ǂ
Test end grade 2	13.39	(.12) ǂ	13.32	(.12) ǂ	13.32	(.12) ǂ	13.32	(.12) ǂ
Test mid grade 3	18.50	(.12) ǂ	18.15	(.13) ǂ	18.15	(.13) ǂ	18.15	(.13) ǂ
Test end grade 3	22.57	(.12) ǂ	22.27	(.13) ǂ	22.27	(.13) ǂ	22.27	(.13) ǂ
Test mid grade 4	25.47	(.14) ǂ	24.94	(.14) ǂ	24.94	(.14) ǂ	24.94	(.14) ǂ
Test end grade 4	29.85	(.14) ǂ	29.34	(.14) ǂ	29.34	(.14) ǂ	29.34	(.14) ǂ
Test mid grade 5	31.60	(.14) ǂ	30.83	(.16) ǂ	30.83	(.16) ǂ	30.83	(.16) ǂ
Test end grade 5	33.09	(.14) ǂ	32.36	(.15) ǂ	32.37	(.15) ǂ	32.37	(.15) ǂ
Test mid grade 6	36.31	(.16) ǂ	35.31	(.18) ǂ	35.31	(.18) ǂ	35.32	(.18) ǂ
Student SES – high			3.14	(.31) ǂ	3.14	(.31) ǂ	3.00	(.32) ǂ
Student SES – low			.46	(.42)	.46	(.42)	.35	(.44)
Student gender (ref=boy)			1.18	(.15) ǂ	1.18	(.15) ǂ	1.18	(.15) ǂ
Student age (months)			.09	(.01) ǂ	.09	(.01) ǂ	.09	(.01) ǂ
Intervention			.79	(.19) ǂ	.89	(.25) ǂ	.61	(.40)
Intervention * StudentSES high							.35	(.22)
Intervention * StudenSES low							.28	(.30)
School Level								
SchoolSize – large			-.06	(.58)	-.09	(.65)	-.09	(.65)
SchoolSize – small			-.99	(.56)*	-1.02	(.60)	-1.02	(.60)
SchoolSESlow			.13	(.60)	.07	(.67)	.04	(.74)
SchoolSEShigh			.85	(.56)	.78	(.64)	.87	(.69)
Trajectory: >1 year spelling (ref = 1st year)					-.62	(.84)	-.65	(.85)
Trajectory: 2nd year spelling (ref = 1st year)					.23	(.59)	.23	(.59)
Intervention * > 1 year spelling					1.13	(.55)	1.18	(.56)
Intervention *2nd year spelling					-.49	(.34) ǂ	-.48	(.37) ǂ
Intervention * SchoolSESlow							.04	(.49)
Intervention * SchoolSEShigh							-.14	(.39)
Variance Components								
Student level								
(Intercept)	33.91		33.90		33.89		33.89	
Clust345	14.30		14.35		14.35		14.35	
Clust678	28.69		28.17		28.15		28.09	

	Model 0	Model 4 Random intervention effect	Model 6 Trajectory* intervention	Model 7 SES* intervention
	Est. (SE)	Est. (SE)	Est. (SE)	Est. (SE)
School level				
(Intercept)	2.19	1.39	1.32	1.33
Intervention		.85	.63	.62
Residual	14.30	14.00	14	14.00
Information Criteria				
AIC	211186.40	210649.4	210649.3	210654.7
BIC	211347.10	210903.2	210936.9	210976.1
-2LogLik	-105574.20	-105295	-105291	-105289
Deviance	211148.40	210589.4	210581.3	210578.7

Note. * *p* < .05; ǂ *p* < .01.

Table 6

Results of hypothesis testing for mathematics and spelling

Hypotheses	Mathematics	Spelling
Hypothesis 1: Student achievement growth will increase as a result of the intervention.	Supported	Supported
Hypothesis 2: School-specific intervention effects will differ across schools.	Supported	Supported
Hypothesis 3: Schools that stick to the same subject for more than one year will have the largest intervention effect.	Supported	Supported
Hypothesis 4: In schools with a high proportion of low-SES students. the intervention effect will be higher	Supported	Not supported
Hypothesis 5: At the student level the intervention effect will be highest for low-SES students.	Partially supported	Not supported

LIST OF FIGURES