Differentiated instruction in a data-based decision-making context

Janke M. Faber, Cees A. W. Glas & Adrie J. Visscher

To cite this article: Janke M. Faber, Cees A. W. Glas & Adrie J. Visscher (2018) Differentiated instruction in a data-based decision-making context, School Effectiveness and School Improvement, 29:1, 43-63, DOI: 10.1080/09243453.2017.1366342

To link to this article: https://doi.org/10.1080/09243453.2017.1366342

© 2017 The Author(s). Published by Informa UK Limited, trading as Taylor & Francis Group.

Published online: 17 Aug 2017.

Article views: 1092

Submit your article to this journal

View related articles

View Crossmark data
Differentiated instruction in a data-based decision-making context

Janke M. Faber, Cees A. W. Glas and Adrie J. Visscher

ABSTRACT
In this study, the relationship between differentiated instruction, as an element of data-based decision making, and student achievement was examined. Classroom observations \( n = 144 \) were used to measure teachers’ differentiated instruction practices and to predict the mathematical achievement of 2nd- and 5th-grade students \( n = 953 \). The analysis of classroom observation data was based on a combination of generalizability theory and item response theory, and student achievement effects were determined by means of multilevel analysis. No significant positive effects were found for differentiated instruction practices. Furthermore, findings showed that students in low-ability groups profited less from differentiated instruction than students in average or high-ability groups. Nevertheless, the findings, data collection, and data-analysis procedures of this study contribute to the study of classroom observation and the measurement of differentiated instruction.

Introduction
Governments expect that data-based or data-driven practices in education will improve student achievement (Mandinach, 2012). Systematic evaluation procedures are encouraged by governments, and the importance of using objective and empirical data for school improvement is emphasized (Kaufman, Graham, Picciano, Popham, & Wiley, 2014; Reis, McCoach, Little, Muller, & Kaniskan, 2011). The idea is that schools systematically collect and organize data, for example, student achievement data, classroom observation data, or parent survey data, in order to represent aspects of school functioning (Schildkamp & Lai, 2013). In this article, we use the definition of Ikemoto and Marsh (2007) for data-based decision making (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” (p. 108).

Educational researchers study how schools implement DBDM procedures. So far, mostly qualitative research findings have indicated which conditions might foster the
implementation of DBDM (Blanc et al., 2010; Levin & Datnow, 2012; Schildkamp & Lai, 2013; Schildkamp, Poortman, & Handelzalts, 2016; Verhaeghe, Vanhoof, Valcke, & Van Petegem, 2010; Wayman, Shaw, & Cho, 2011; Wayman, Stringfield, & Yakimowski, 2004), and which conditions hinder DBDM procedures in schools (Ehren & Swanborn, 2012; Schildkamp & Lai, 2013). Research results pertaining to the actual effect of DBDM on student achievement vary (Marsh, 2012). Some researchers found no effect at all (Cordray, Pion, Brandt, Molefe, & Toby, 2012), others only a significant effect for specific groups of students (May & Robinson, 2007), and again others an overall significant improvement of student achievement (Carlson, Borman, & Robinson, 2011; Konstantopoulos, Miller, & van der Ploeg, 2013; Van Geel, Keuning, Visscher, & Fox, 2016; Van Kuijk, Deunk, Bosker, & Ritzema, 2016). Although each of these studies focused on DBDM practices, the interventions varied (Marsh, 2012). In the study by Van Kuijk et al. (2016), teachers, in addition to learning about DBDM, also learned new instructional skills and knowledge for teaching reading comprehension, making it difficult to interpret the contribution of DBDM to the effect found in the study. The studies of Cordray et al. (2012) and May and Robinson (2007) both focused on the use of assessment data for DBDM. In the first study, teachers used adaptive standardized tests administrated three to four times a year, while the data fed back to teachers in the second study were based on a single state test. Furthermore, in some studies the whole school team including school principals and academic coaches were involved (Van Geel et al., 2016), while in others only one teacher per school participated (Van der Scheer, 2016). In sum, we still know little about how DBDM can be best used to improve student achievement. Interventions and their effects vary, and, as DBDM quite often is a component of an intervention package, it is unclear precisely which intervention component caused the observed student-achievement effects.

DBDM seems a promising approach for school development and instruction improvement (Carlson et al., 2011; Konstantopoulos et al., 2013; Mandinach, 2012; Van Kuijk et al., 2016); however, knowledge on how DBDM affects achievement is lacking. The aim of this study was to contribute to filling that knowledge gap. Differentiated, targeted, or adapted instruction is a frequently mentioned aspect of classroom DBDM (Black & Wiliam, 1998; Hamilton et al., 2009), as instructional strategies are supposed to be grounded on student data, and adapted in line with differences in learning needs between students. We therefore examined the relationship between differentiated instruction and student achievement in a DBDM context. The Dutch Ministry of Education encourages teachers, principals, and administrators to develop their abilities to analyze, interpret, and use student data to guide improvement (Ministerie van Onderwijs, Cultuur en Wetenschap, 2007). Furthermore, the Ministry facilitated several DBDM projects, among others, the Focus intervention (Staman, Visscher, & Luyten, 2014; Van Geel et al., 2016). In the Focus intervention, whole school teams were trained in DBDM. All teachers selected for this study participated in the Focus intervention. Before we explain the Focus intervention, we first characterize DBDM and differentiated instruction.

**Theoretical framework**

DBDM has become an important area of interest within the field of educational research (Hamilton et al., 2009; Mandinach, 2012; Schildkamp, Ehren, & Lai, 2012). Although the
concept of using data to guide teacher and school improvement is not new (Mandinach, 2012), as a result of DBDM data are supposed to be collected and used by schools in a more systematic and cyclic way (Van Geel et al., 2016). Data have to be identified, collected, analyzed, and interpreted before actions can be taken (Ikemoto & Marsh, 2007; Schildkamp & Kuiper, 2010; Van der Kleij, Vermeulen, Schildkamp, & Eggen, 2015). In Figure 1, the DBDM cycle used in the Focus intervention is presented. This DBDM cycle starts with the analysis and interpretation of student-achievement data, such that teachers collect information on students’ progress. This information gives teachers an indication of the extent to which their instruction matches students’ needs or how effective their teaching was (Hattie & Timperley, 2007). Based on this, teachers can set realistic performance goals for students. In the case of clearly formulated and challenging goals, teachers can collect feedback that is more targeted at goal accomplishment and can better examine the results of a new instruction strategy (Locke & Latham, 2002). In the third step, determining the instructional strategy, teachers are supposed to choose teaching strategies matching students’ needs, based on the first and second steps. Teachers implement their planned teaching strategies in the classroom in the fourth step, after which the cycle starts again with the analysis of the effects of the strategies implemented.

In DBDM, the four components of the cycle in Figure 1 are related and connected. The focus in this study, however, is on the last component: executing an instructional strategy. The impact of a planned instructional strategy will probably be more positive if the first three components have been carried out successfully. As the strategies are based on data concerning the performance of individual students, instructional strategies are ideally more individualized since learning needs of students differ. When DBDM leads to an instructional strategy which adapts to students’ learning needs, and when this strategy is executed by teachers, student achievement might improve. However, teachers require differentiation skills to execute such an instructional strategy (Mandinach & Gummer, 2013).

**Differentiated instruction**

A well-known definition of differentiated instruction (DI) was provided by Tomlinson et al. (2003): “Teachers proactively modify curricula, teaching methods, resources,
learning activities, and student products to address the diverse needs of individual students and small groups of students to maximize the learning opportunity for each student in a classroom’ (p. 121). Tomlinson et al. also state that teachers should proactively modify teaching to address a broad range of learners’ readiness levels, interests, and modes of learning. This definition is rather broad, and Roy, Guay, and Valois (2013) argue that DI is a varied and adapted teaching approach to match students’ abilities, or students’ readiness levels. Tomlinson et al. and Roy et al. conceptualized DI differently. According to Tomlinson et al., DI is proactive. Teachers plan a lesson beforehand to address learner variance. Second, teachers work with flexible and small teaching-learning groups. Since students differ in their readiness, interests, and modes of learning, it is important to group them in a variety of ways. Third, teachers provide variation in learning materials, instruction time, and pace in their classrooms. A final important DI characteristic, according to Tomlinson et al., is the learner-centered classroom. In a learner-centered classroom, teachers focus on the needs of all students and as a result use a wide variety of instruction strategies and practices. Roy et al. used two distinct DI components in their conceptualization of DI: instructional adaptation and academic progress monitoring. The first component, instructional adaptation, overlaps to a large extent with Tomlinson et al.’s third DI characteristic (teachers provide for variation in learning materials, instruction time, and pace in their classrooms). In addition to this, Roy et al. argue that modifying the goals and expectations for students with difficulties is another important aspect of instructional adaptation. An important difference between the two conceptualizations is that Roy et al. place greater emphasis on the need of academic progress monitoring for instructional adaptations, while Tomlinson et al. only mention that DI should be proactive. According to Roy et al., academic progress monitoring can be measured by the degree to which teachers evaluate the effects of their teaching adaptations, the degree to which teachers analyze data about student progress, their use of student data for instructional decisions, and finally whether teachers frequently assess low-performing students’ rates of improvement.

Two important DI characteristics emerge from these conceptualizations. First, DI is planned, and instructional decisions should be based on the analysis of student data. Second, what makes DI observable in the classroom is the variation in learning goals, instruction content, instruction time, assignments, and learning materials aimed at addressing varying learning needs. In the present study, we tested whether these DI characteristics explain student achievement.

Within-class grouping

In many cases, providing DI for each individual student is unrealistic, as it is in Dutch classrooms, for instance, accommodating 23 students on average (Inspectie van het Onderwijs [Inspectorate of Education], 2015). Small instruction groups are therefore often used to organize DI and to vary learning goals, instruction content, instruction time, assignments, and learning materials within relatively large classrooms. Meta-analysis findings reveal that small-group instruction can have a positive effect on student achievement (Lou et al., 1996). The effects are, however, influenced by how the groups are composed. In homogeneous groups, students of the same ability level are grouped in one group (ability
grouping), whereas heterogeneous groups include students from different ability levels. Low-ability students seem to learn more in heterogeneous groups, average-ability students learn more in homogeneous ability groups, and for high-ability students the grouping composition does not make much of a difference (Lou et al., 1996; Saleh, Lazonder, & De Jong, 2005). School characteristics also influence the effects. Ability grouping has a positive effect in schools with a homogeneous student population, or a high-socioeconomic status (SES) student population. In schools with a low-SES student population or a heterogeneous population, ability-grouping effects prove to be negative for low-ability students (Nomi, 2009). These findings seem remarkable for two reasons. First, assuming that variation in instruction is necessary since learning needs differ (Smit & Humpert, 2012; Tomlinson et al., 2003), one would expect that ability grouping is required more in classrooms in which learning needs differ than in classrooms in which learning needs are more similar (as in the case of a homogeneous student population). Second, assuming that regular classroom instruction is mostly tailored to the average ability level, one would expect that especially low-ability and high-ability students profit from ability grouping (instead of average-ability students).

There are several explanations for these findings. The first may be that students learn by giving and receiving explanations. Following this assumption, low-ability students learn in heterogeneous groups by receiving explanations from peers. Average-ability students act to a greater extent as explanation receivers and providers in homogeneous groups, and high-ability students learn in heterogeneous groups by being a tutor (Lou et al., 1996; Saleh et al., 2005). These explanations relate to assumptions about the effectiveness of student collaboration. Other explanations refer more specifically to DI. These explanations relate to the negative effects of DI on low-ability students. Teachers might lower their expectations for these students (Campbell, 2014; Wiliam & Bartholomew, 2004), and more time may be spent on behavior management than on instruction (Wiliam & Bartholomew, 2004). The time spent by a teacher on a specific group requires self-regulation skills from those students who are not placed in that group, and especially low-ability students might find this difficult (Hong, Corter, Hong, & Pelletier, 2012). Teachers might be better equipped with curriculum materials and pedagogical skills tailored at students in the middle of the ability range, which might explain the positive effects on average-ability students (Hong et al., 2012).

The effectiveness of ability grouping will strongly depend on how teachers implement DI and how they organize ability grouping. It is important that teachers base the composition of their instruction groups on various data resources, and not just on achievement data (Houtveen, Booij, De Jong, & Van de Griff, 1999; Lou et al., 1996). Other characteristics of effective grouping practices are that the groups are based on the skill being taught, that the grouping composition is flexible (Deunk, Doolaard, Smale-Jacobse, & Bosker, 2015; Slavin, 1987), and that learning materials (Lou et al., 1996) and learning time (Slavin, 1987) vary between ability groups. Although ability grouping has been studied frequently (Deunk et al., 2015; Lou et al., 1996; Slavin, 1987), more research is needed on the effect of using within-class homogeneous ability grouping on student achievement.

**Hypotheses**

Based on the literature, we tested the following hypotheses:
Hypothesis 1: Student outcomes are higher in the classrooms of teachers who differentiate their instruction more (observable differentiation).

Hypothesis 2: Student outcomes are higher in the classrooms of teachers who preplanned DI more (planned differentiation).

Hypothesis 3: Students from different ability groups do not benefit to the same degree from a teacher who differentiates his/her instruction.

Method

In this section, we describe how a pretest-posttest observational study was used to test our hypotheses. Furthermore, we describe the nature of our sample, instruments, data-collection procedures, and the data-analysis techniques used. First, the features of the Focus intervention and the context in which the research was carried out are described. Descriptive statistics are presented in Table 1.

The focus intervention

In the intervention, entire primary school teams were trained in analyzing student data, formulating performance student learning goals, formulating instructional strategies that match students’ needs, and providing targeted instruction (Staman et al., 2014; Van Geel et al., 2016). More specifically, teachers learned how to analyze Cito assessment results (Cito is the Dutch institute for test development; see the section Standardized assessments for a description of these assessments) by using a student monitoring system. Teachers learned how to design an instructional plan (which included teachers’ differentiated instruction decisions) twice a year based on those assessment results. Furthermore, teachers learned how to assign students more systematically, and in a

Table 1. Descriptive statistics for students and teachers: means, standard deviations, and total number of respondents.

<table>
<thead>
<tr>
<th></th>
<th>Grade 2</th>
<th>Grade 5</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>N (missing)</td>
<td>Mean (SD)</td>
</tr>
<tr>
<td>Teachers</td>
<td>25</td>
<td></td>
</tr>
<tr>
<td>Multigrade group</td>
<td>32%</td>
<td></td>
</tr>
<tr>
<td>Students</td>
<td>466</td>
<td></td>
</tr>
<tr>
<td>Student weight low</td>
<td>18.5% (1.9%)</td>
<td></td>
</tr>
<tr>
<td>Ability group</td>
<td></td>
<td></td>
</tr>
<tr>
<td>– Low</td>
<td>107</td>
<td></td>
</tr>
<tr>
<td>– Average</td>
<td>185</td>
<td></td>
</tr>
<tr>
<td>– High</td>
<td>168</td>
<td></td>
</tr>
<tr>
<td>– Missing</td>
<td>6</td>
<td></td>
</tr>
<tr>
<td>Cito May/June 2013</td>
<td>443 (4.9%)</td>
<td>46.24 (15.60)</td>
</tr>
<tr>
<td>Cito May/June 2014</td>
<td>454 (2.6%)</td>
<td>64.80 (14.38)</td>
</tr>
<tr>
<td>Differentiated instructiona</td>
<td>216 (3)</td>
<td>2.54 (0.70)</td>
</tr>
<tr>
<td>Planned Di b</td>
<td>46 (4)</td>
<td>17.7 (5.92)</td>
</tr>
</tbody>
</table>

aThe numbers reflect the number of completed ICALT observation forms. bThe maximum score for the instructional plan checklist was 43.
data-based way, to ability groups by using Cito assessments. Students are categorized into one of five performance categories (A: highest; E: lowest) based on the results of the Cito standardized test. However, the assignment of students to class ability groups depends on how the students of that specific class performed on the test. Relative to the other students in the same class, the best performing students in a class are assigned to the high-ability group by their teacher, the lowest performing students of that class are assigned to the low-ability group, and all other students are assigned to the average-ability group. Ability groups are composed based on students’ test results and the choices teachers make based on these results. The use of Cito standardized assessment data was the main focus of the intervention; however, teachers were also stimulated to use other student data (e.g., their classroom observations of students) for making instruction decisions, and not to concentrate only on the results of standardized assessments.

The instructional plan included teachers’ differentiated instructional strategies. Instructional plans had to include three within-class homogeneous ability groups, an instructional strategy and learning goals for each of these three groups, just as an additional instructional approach for one or a few individual students with specific learning needs. In the intervention, teachers divided students over three homogeneous ability groups: in the “average” group students receive regular instruction, in the “low” group students require some additional instruction, and in the “highest” group students only need a brief part of the regular instruction (Van der Scheer, 2016). An instructional plan format was used to ensure all elements of the instructional plan were covered by teachers.

Trainers of the Focus intervention delivered between five and seven school team meetings, and attended two additional meetings with the school principal in one school year (the duration of the whole intervention was 2 years). During the first part of the intervention, the skills and knowledge required for executing the DBDM cycle (Figure 1) were trained. Teachers also received feedback on their teaching based on classroom observations. The meetings with school principals were meant to support principals in motivating teachers for DBDM. For a more detailed description of the Focus intervention, see Keuning, Van Geel, Visscher, Fox, and Moolenaar (2016), Staman et al. (2014), and Van Geel et al. (2016).

**Sample**

A fraction of all schools involved in the intervention were selected for this study. We contacted schools by email and also visited several schools. School principals and teachers were informed about the purpose and design of the study, and the planned classroom observations. Trainers of the intervention had rated each of their schools as a “weak”, “average”, or “strong” DBDM school for the school selection procedure of this study. Our goal was to include 10 schools from each of these three categories to ensure variation in DBDM practices between schools; however, not all invited schools agreed to participate. Twenty-six schools agreed to participate: 7 “weak”, 9 “average”, and 10 “strong” DBDM schools. Most of these schools had already finished the whole intervention, and 6 schools were in the last half year of the intervention. Only second-grade teachers (7/8-year-old students), and fifth-grade teachers (10/11-year-old students)
participated. Second and fifth grades were selected to ensure that teachers and students from the lower and the upper grades participated. During the course of the study, one teacher refused to be videotaped. As a result, the final sample included 26 primary schools, 51 teachers, and 953 students (Table 1).

Nineteen percent of the students had a student weight (in The Netherlands, primary schools receive extra funding for these students, and a child belongs to this category of students if neither of both parents attained a higher qualification than lower vocational education), which, compared with a national average of 11%, is rather high (Centraal Bureau voor de Statistiek [CBS], 2014). Thirty-two percent of the second-grade classes and 35% of the fifth-grade classes were multigrade classes. In multigrade classes, two different age groups are combined. Eighty-four percent of the teachers in second grade and 65% in fifth grade did not teach their class alone, but had a teacher-colleague teaching the same class during part of the week. Of each classroom, one teacher participated in the study (e.g., the teacher who taught mathematics most in that classroom). Since more than two thirds of the primary education teachers in The Netherlands have a part-time teaching job, these numbers are not extreme (Inspectie van het Onderwijs, 2012).

**Instruments and procedures**

***ICALT***

We used the ICALT (International Comparative Analysis of Learning and Teaching) classroom observation instrument to measure teachers’ observable DI (Hypothesis 1). The ICALT is based on research literature on teaching effectiveness and was validated in international comparative studies. Findings indicated reliable and valid measurements of the six aspects of teaching included in ICALT (Van de Grift, 2007, 2014). Each aspect of teaching included items rated on a 4-point Likert scale ranging from *mainly weak* to *mainly strong*. The following items were used:

- The teacher evaluates whether the lesson objectives have been achieved at the end of the lesson.
- The teacher offers extra learning and instruction time to struggling learners.
- The teacher adapts his/her instructional activities to relevant differences between students.
- The teacher adapts the assignments to relevant differences between students.

Since the first item lowered the scale’s internal consistency, this item was excluded from the further analyses. Cronbach’s alpha for the remaining items was $\alpha = 0.73$.

Generalizability studies have shown that reliable teacher ratings, on the basis of lesson observations, require two or more observed lessons per teacher in combination with two or more raters (Hill, Charalambous, & Kraft, 2012). Therefore, three mathematics lessons of each teacher were taped and afterwards rated by three trained raters. Recordings were spread over a period ranging from November 2013 to May 2014. Teachers were asked to give an entire regular mathematics lesson lasting between 45 and 60 min. We used the IRIS Connect toolkit for recording the lessons: A system with two mobile devices simultaneously recorded the teacher and the students. Recordings were uploaded to a secured, online environment. Principals of participating schools decided if and how parents were
requested for permission, and children of parents who did not provide permission were not recorded. Raters followed a 3-day training course, and rated six observations independently of each other during this course. Afterwards, rater variation was discussed and rating guidelines were developed to maximize consensus between raters. Raters agreed to watch all taped lessons randomly, to prevent order-based bias. Due to changes in teacher teams (i.e., maternity leave, illness) or planning problems, not all participating teachers were recorded three times. A total of 144 lessons were recorded (nine teachers with one lesson missing) and scored by each rater. Of nine teachers, their DI scores were computed with data from two lesson observations instead of three.

**Instructional plan checklist**

Teachers are supposed to plan DI in advance in order to address learner variance (Tomlinson et al., 2003). We collected teachers’ instructional plans to measure their planned DI (Hypothesis 2). Teachers in the intervention learned to develop two instructional plans based on the standardized assessments in January/February and May/June (see section Standardized assessments). Of each classroom, all mathematics instructional plans covering the same period as the observations were collected. We collected 46 second-grade plans (4 missing) and 43 fifth-grade plans (9 missing). To measure planned DI, a checklist was developed to evaluate the instructional plans. Prototypes of the checklist were tested by Focus trainers to make sure all elements of the instructional plan as learned during the intervention were included. Since teachers learned to use a format, teachers’ instructional plans were very similar. The checklist consisted of 43 items to measure the degree to which teachers vary the following three topics between ability groups:

1. instruction (e.g., learning materials, learning pace);
2. learning goals (e.g., the specification of a percentage of minimally required correct answers on the mathematics test);
3. evaluation (e.g., the specification of follow-up actions in case learning goals were not yet accomplished).

Furthermore, the degree to which teachers specified the instruction for students with specific learning needs was measured with this checklist (see Appendix 1). Two raters evaluated all instructional plans and scored each item in the checklist as “present” or “not present” (intraclass correlation coefficient [ICC] = 0.63, two-way mixed model, absolute agreement, IBM SPSS Statistics Version 22). Aggregated total checklist scores were included in the analyses (i.e., the average of four checklist scores: two instructional plans, each scored by two raters).

**Standardized assessments**

Students’ results on the Cito standardized mathematic tests were used for the dependent variable, student achievement. Most Dutch schools use these tests for the whole primary school period (all grades). The results from different grades can be placed on one and the same ability scale. The Cito mathematic tests measure three different domains: (a) arithmetic; (b) proportions, fractions, and percentages; and (c) geometry, time, and money calculations. Students take the tests twice a year: in January or
February, and in May or June. For the dependent variable, the standardized test results of May or June 2014 were used, whereas the standardized results of the May or June 2013 test were used as a covariate (pretest). The percentage of missing data was 7.3% for the pretest and 2.1% for the posttest. Within-group regression, using the group mean and the other test score as predictors was used as an imputation method. So, a regression model was estimated based on all complete cases, using the group means and all available scores, and based on this regression model, the missing values were estimated and imputed.

**Analysis**

Since students are nested within teachers’ classes, a multilevel model was used to test the hypothesis regarding the relation between students’ achievement scores and teachers’ differentiated instruction skills. DI and planned DI are classroom or group-level variables; all other variables included in the analysis were measured at the student level. The Level 1 model is given by:

\[ Y_{ij} = \beta_{0j} + \beta_1X_{1ij} + \beta_2X_{2ij} + \beta_3X_{3ij} + \beta_4X_{4ij} + \beta_5X_{5ij} + R_{ij}, \]

where \( Y_{ij} \) represents the standardized posttest score for student \( i \) of teacher \( j \). The five covariates are the 2013 score, Gender, the Student Weight defined in the Sample section, and two dummy codes for Ability Group (high and low), respectively. Note that \( R_{ij} \) is a residual and that the intercept \( \beta_{0j} \) is random over teachers. The model for this random coefficient is given by:

\[ \beta_{0j} = \gamma_0 + \gamma_1\theta_j + \gamma_2Z_{2j} + \gamma_3Z_{3j} + U_{0j}, \]

where \( Z_{2j} \) and \( Z_{3j} \) are teacher-level variables Grade and Aggregated Planned DI scores, respectively. \( U_{0j} \) is the Level 2 residual. Finally, \( \theta_j \) stands for the latent variable Differentiated Instruction (DI). So \( U_{0j} \) and \( R_{ij} \) are the deviations from the average score for teachers and the deviations of students from the teachers’ average, given the covariates at the two levels (Luyten & Sammons, 2010; Snijders & Bosker, 1999).

This latent variable was analyzed with a combined item response theory (IRT) model (Lord, 1980), and a generalizability theory (GT) model (Brennan, 1992). All teachers were rated by raters (indexed \( r \)) using the same items (indexed \( k \)) at all time points (indexed \( t \)). The items were scored on a Likert-point scale with categories indexed \( h \) (\( h = 0, \ldots, 3 \)). The IRT model was the generalized partial credit model. The function of the IRT model was to map the discrete item responses to a continuous overall measure, that is, to a latent variable. For teacher \( j \) at a time point \( k \), the probability of a score in category \( h \) of item \( k \) denoted by \( U_{jtrk} = h \) is given by:

\[ P(U_{jtrk} = h) = \frac{\exp(ha_k\theta_{jtr} - \delta_{kh})}{1 + \sum_{g=1}^{3}\exp(ga_k\theta_{jtr} - \delta_{kh})}, \quad h = 0, 1, 2, 3, \]

where \( \theta_{jtr} \) is the position on the latent variable of teacher \( j \) on time point \( t \) as judged by rater \( r \). Note that the responses were recoded from a scale that ran from 1 to 4, to a scale from 0 to 3. Further, in Equation (3), \( a_k \) and \( \delta_{kh} \) are the parameters of item \( k \); \( a_k \) is an item discrimination parameter that gauges the relation between the observed score and
the latent scale, and $\delta_{kh}$ ($h = 1, \ldots, 3$) represent locations on the latent scale, that is, they gauge the salience of the item. A GT model was imposed on the latent variables $\theta_{jtr}$, that is:

$$\theta_{jtr} = \theta_j + \tau_{1t} + \tau_{2r} + \tau_{3jt} + \tau_{4jr} + \tau_{5tr} + \varepsilon_{jtr}$$  (4)

Note that $\theta_j$ is the main effect for the teacher, $\tau_{1t}$ and $\tau_{2r}$ are the main effects for the time points and raters, and the other terms are the two-way interaction effects and a residual.

The complete model, given by the Equations (1) to (4), was estimated in a Bayesian framework using OpenBUGS (Version 3.2.3. rev. 2012). We built the model from an empty model with only the standardized posttest scores to the final model that also included the interaction effect between the ability-group division measured at the student level, and the latent DI scores measured at the group level. Besides testing the hypotheses using the multilevel model, the rater reliability was also of interest. This reliability was estimated using the variance decomposition of the GT model; this procedure is generally known as a generalizability study (Brennan, 1992). The reliability coefficient is given by:

$$\rho = \frac{\sigma_j^2}{\sigma_j^2 + \sigma_{jt}^2 + \sigma_{jr}^2 + \sigma_e^2}$$  (5)

where $\sigma_j^2$ is the variance of teachers, $\sigma_{jt}^2$ and $\sigma_{jr}^2$ are the variances associated with the interaction of teachers with time points and raters, and $\sigma_e^2$ is the error variance.

A Bayesian estimation as implemented in OpenBugs is an iterative process based on a Markov chain Monte Carlo (MCMC) method. All parameters were estimated with 4,000 burn-in iterations and 16,000 effective iterations. The estimation procedures were repeated several times, and the Monte Carlo standard error for all parameters was always well below 5%. Standard prior distributions were used for all parameters; that is, means and regression parameters had normal priors with mean zero and low precisions (0.25), the inverses of variance parameters had uninformative Gamma distributions, the item discrimination parameters had normal priors with a mean of 1.0, and a variance of 1.0, truncated to the positive domain, and the item location parameters had standard normal priors. Furthermore, Pearson’s (frequentist) correlation coefficients were computed using aggregated DI scores, aggregated planned DI scores, and aggregated and disaggregated student pretest and posttest scores.

## Results

We first present the results of the estimation of the GT model to assess rater reliability and obtain the latent teacher variables needed for the multilevel model. Table 2 presents the variance components in classroom observations data for the DI scale needed to compute rater reliability. Teachers’ variance was set to 1.00 to identify the scale. It is important to note that analyses with latent variables require an identification of the location and the scale. The scale is identified using the teacher’s variance. This is done for convenience; that is, it is the largest variance, but any other choice would have produced exactly the same results since all variances would have been multiplied by the
same constant, and their ratio would not change. The percentages in the table represent the percentages of variance explained by the specific component compared to the total variance. From the percentages, it can be seen that almost 37% of the variance is explained by differences between teachers. This result indicated that most of the variance in the scale was due to differences between teachers. Furthermore, the differences between the three observation moments explained 19.33% of the variance. Compared to the other variance components, this is high. The proportion of variance explained by differences between raters was much smaller (7.67%), indicating that raters predominantly agreed upon the rankings of the teachers. This consistency in ratings is also reflected by the reliability coefficient, this coefficient was 0.83 (SD = 0.02) for DI. The influence of different observation moments is reflected by the interaction between teachers and time moments, which shows the extent to which the ordering of teachers over different time moments explains variance. This interaction explains 12.72% of the variance.

Table 3 shows the Pearson’s correlation coefficient between explanatory variables and the dependent variable. Aggregated teachers’ DI and planned DI scores were used for computing the correlations. DI has significant positive correlations with the pretest scores (r DI pretest = 0.19, p < 0.01), and the posttest scores (r DI posttest = 0.17, p < 0.01). Interestingly, the correlations between planned DI and achievement were negative and significant (r pretest = −0.13, p < 0.01 and r posttest = −0.15, p < 0.01). Also, the small and nonsignificant correlation between DI and planned DI (r = 0.02, ns) is remarkable, since it seems reasonable that planning DI is related to executing DI practices in the classroom. The same patterns are found in the correlations with

<table>
<thead>
<tr>
<th>Table 2. Variance components observation data.</th>
</tr>
</thead>
<tbody>
<tr>
<td>Component</td>
</tr>
<tr>
<td>σ_j</td>
</tr>
<tr>
<td>σ_t</td>
</tr>
<tr>
<td>σ_r</td>
</tr>
<tr>
<td>σ_{jt}</td>
</tr>
<tr>
<td>σ_{fr}</td>
</tr>
<tr>
<td>σ_{tr}</td>
</tr>
<tr>
<td>σ_e</td>
</tr>
<tr>
<td>Total</td>
</tr>
</tbody>
</table>

Note: SD stands for posterior standard deviation, which can be interpreted as a standard error.

<table>
<thead>
<tr>
<th>Table 3. Correlations.</th>
</tr>
</thead>
<tbody>
<tr>
<td>DI</td>
</tr>
<tr>
<td>DI</td>
</tr>
<tr>
<td></td>
</tr>
<tr>
<td>Planned DI</td>
</tr>
<tr>
<td></td>
</tr>
<tr>
<td>Pretest</td>
</tr>
<tr>
<td>Posttest</td>
</tr>
</tbody>
</table>

Note: Scores below diagonal correlations with aggregated (classroom level) student pre- and posttest scores.

*Correlation is significant at the 0.01 level (2-tailed).
aggregated student test scores (classroom level); only, these correlations are somewhat higher (with regard to DI) or lower (with regard to planned DI).

The results of the multilevel analysis, as shown in Table 4, show that in the empty model a large proportion of the variance is group-level variance (variance = 0.87, SD = 0.19), leading to a high ICC of 0.69. This high proportion of group-level variance is caused by combining the assessment results from two different grades (second- and fifth-grade students) into the model. The ICC drops in Model 1 to 0.13, after including student grade and student pretest scores. This result indicates that, as expected, most variance is due to differences between students and to a lesser degree due to differences between groups.

Covariates were included in Model 1. Significant covariates were gender (β = −0.08, p < 0.05; significance indicates that the value zero is outside the 1% Bayesian credibility region), grade (β = 0.39, p < 0.05), and pretest (β = 0.76, p < 0.05). Similar effects for gender were also found in the next models, which indicates that boys’ mathematic achievement was significantly higher than girls’ mathematic achievement. As expected, the achievements of students in second grade were lower than the achievements of students in fifth grade. Furthermore, students with a high pretest scored higher on the posttest than students with a lower pretest score. No significant effects of student weight were found.

Explanatory variables were included in subsequent models. In Model 2, standardized planned DI scores were added. No significant positive effects were found for planned DI. This finding does not support Hypothesis 2: Students’ outcomes are higher in the classrooms of teachers who planned DI more. In Model 3, the latent DI observation scores were included. Again, no significant positive effects were found. This finding does not support Hypothesis 1: Students’ outcomes are higher in the classrooms of teachers who differentiate their instruction more.

Table 4. Multilevel analysis results.

<table>
<thead>
<tr>
<th>Predictors</th>
<th>Model 0</th>
<th>Model 1</th>
<th>Model 2</th>
<th>Model 3</th>
<th>Model 4</th>
<th>Model 5</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Mean</td>
<td>SD</td>
<td>Mean</td>
<td>SD</td>
<td>Mean</td>
<td>SD</td>
</tr>
<tr>
<td>Fixed</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Intercept</td>
<td>−0.13</td>
<td>0.13</td>
<td>1.01*</td>
<td>0.20</td>
<td>1.05*</td>
<td>0.19</td>
</tr>
<tr>
<td>Gender (1 = girl)</td>
<td>−0.08*</td>
<td>0.03</td>
<td>−0.08*</td>
<td>0.03</td>
<td>−0.08*</td>
<td>0.03</td>
</tr>
<tr>
<td>Student weight (0 = no weight)</td>
<td>−0.02</td>
<td>0.04</td>
<td>−0.02</td>
<td>0.04</td>
<td>−0.02</td>
<td>0.04</td>
</tr>
<tr>
<td>Grade (0 = second)</td>
<td>0.39*</td>
<td>0.08</td>
<td>0.38*</td>
<td>0.08</td>
<td>0.38*</td>
<td>0.08</td>
</tr>
<tr>
<td>Pretest</td>
<td>0.76*</td>
<td>0.03</td>
<td>0.75*</td>
<td>0.03</td>
<td>0.76*</td>
<td>0.03</td>
</tr>
<tr>
<td>Planned DI</td>
<td>−0.05</td>
<td>0.03</td>
<td>−0.05</td>
<td>0.03</td>
<td>−0.05</td>
<td>0.03</td>
</tr>
<tr>
<td>Differentiated instruction (DI)</td>
<td>−0.02</td>
<td>0.05</td>
<td>0.00</td>
<td>0.05</td>
<td>0.05</td>
<td>0.06</td>
</tr>
<tr>
<td>Ability group</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Low</td>
<td>−0.24*</td>
<td>0.04</td>
<td>−0.22*</td>
<td>0.05</td>
<td></td>
<td></td>
</tr>
<tr>
<td>High</td>
<td>0.41*</td>
<td>0.04</td>
<td>0.41*</td>
<td>0.04</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Ability group * DI</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Low * DI</td>
<td>−0.20*</td>
<td>0.09</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>High * DI</td>
<td>−0.03</td>
<td>0.06</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Random</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Variance student level</td>
<td>0.37</td>
<td>0.02</td>
<td>0.20</td>
<td>0.01</td>
<td>0.20</td>
<td>0.01</td>
</tr>
<tr>
<td>Variance group level</td>
<td>0.87</td>
<td>0.19</td>
<td>0.03</td>
<td>0.01</td>
<td>0.03</td>
<td>0.01</td>
</tr>
<tr>
<td>ICC</td>
<td>0.69</td>
<td>0.04</td>
<td>0.13</td>
<td>0.03</td>
<td>0.12</td>
<td>0.03</td>
</tr>
</tbody>
</table>

Note: SD stands for posterior standard deviation, which can be interpreted as a standard error.
Ability-group and interaction effects were added in Models 4 and 5. As expected, students in low-ability groups had significant lower achievement scores than students in average-ability groups ($\beta = -0.24, p < 0.05$). Furthermore, students in high-ability groups had significantly higher achievement scores than students in average-ability groups ($\beta = 0.41, p < 0.05$). In Model 5, the interaction between ability group and DI was included. From the results, it follows that students in low-ability groups whose teachers had high DI scores had significantly lower posttest scores than students in average-ability and high-ability groups with teachers who also had high DI scores ($\beta = -0.22, p < 0.05$). This finding supports Hypothesis 3: Students from different ability groups do not benefit equally from teachers who differentiate their instruction. Our results suggest that, compared with students in high- or average-ability groups, students in low-ability groups profit significantly less from a teacher with high DI observation scores. No significant results were found for the interaction between high-ability-group students and DI.

Discussion and conclusions

DBDM-intervention effects on student achievement have been examined in several research projects (Carlson et al., 2011; Cordray et al., 2012; Konstantopoulos et al., 2013; May & Robinson, 2007; Van Geel et al., 2016; Van Kuijk et al., 2016); however, our knowledge of how DBDM affects student achievement is still very limited. The purpose of the present study was to investigate the relationship between DI and student achievement in a DBDM context.

First, the findings of the generalizability study showed that, even though most variance was explained by differences between teachers, there was much variability between the lessons of the same teacher. These observation time effects were also found in a study by Praetorius, Pauli, Reusser, Rakoczy, and Klieme (2014), and such findings indicate that more research is needed on how valid and representative teacher observation scores can be obtained. Furthermore, our findings indicated that students from different ability groups do not profit from DI to the same extent. This finding is in line with previous research: Ability grouping can have a negative impact on the achievement of students in low-ability groups, ability grouping is effective for students in average-ability groups, and ability grouping has no impact on the achievement of students in high-ability groups (Lou et al., 1996; Saleh et al., 2005). In future research, it would be worth investigating whether lower teacher expectations, less stimulating learning materials, and a lack of self-regulation skills among low-performing students (Campbell, 2014; Hong et al., 2012; Nomi, 2009; Wiliam & Bartholomew, 2004) could explain the negative impact of DI on the achievement of students in low-ability groups. Furthermore, we expected that students taught by teachers who differentiate their instruction more, or by teachers who plan DI more, have higher student achievement levels. No such positive effects were found. A reverse causality between DI and student achievement (i.e., DI practices are executed more in classrooms with many low-performing students and a very diverse student population) might be an explanation for this finding (De Neve & Devos, 2016; Nomi, 2009). Another explanation might be the impact of DI on noncognitive outcomes such as students’ feelings of competence (Carver & Scheier, 1990). Especially for students in low-ability groups, there might have been an
impact on noncognitive outcomes, and consequently on student achievement. Also, these findings may suggest that planning differentiation strategies in advance should always be combined with responsive ad hoc classroom differentiation practices. It may be that a balance between preplanned instruction and responsive teaching is most effective (Sawyer, 2004). In future studies, such effects should be studied to explain better how DBDM affects student achievement.

Even though no relation between DI (as a preplanned teaching approach) and achievement was found in the present study, the findings still contribute to the data collection procedures that can be used to measure DI. DI cannot be measured with observations alone, as it is necessary to know the rationale behind the differentiation approaches observed (Allen, Matthews, & Parsons, 2013). Variation in learning material, instruction time, and assignments between students is easily observed by means of classroom observations; however, it is difficult for the rater to judge whether this observed differentiation in instructional activities matches the instructional needs of students. The findings of this study indicate that teachers’ differentiation practices in classrooms and their preplanned differentiation practices on paper are not always related. Therefore, we recommend that in future DI effectiveness studies other measures aimed at determining the rationale behind differentiation activities will be used to assess the fit between DI and actual instructional needs. For example, to determine the rationale behind the observed DI and examine whether the observed DI indeed matched students’ needs, students and teachers could be interviewed immediately after classroom observations. Research findings have already indicated that students are able to judge teachers’ behavior, as students’ perceptions of teacher behavior proved to be good predictors for student outcomes (Maulana, Helms-Lorenz, & Van de Grift, 2015).

Furthermore, the data analysis procedures used in this study contribute to the existing knowledge base. They solve many problems of studies with observations made by multiple raters, on multiple time points, using itemized observation instruments. First of all, generalizability as such is a tool for disentangling the variance components in classroom observation data, and for making an informed choice regarding the indices of reliability and agreement that best fit the purpose of the observation. However, traditionally, the model is imposed on directly observed sum scores. This ignores measurement error at the item level. Using an IRT model as a measurement error model can reduce bias, for instance, caused by floor and ceiling effects. Furthermore, IRT models are much more flexible in handling missing item responses and better suited for optimization of measurement designs (e.g., by optimal item selection). Another aspect of the innovative approach concerns the Bayesian framework combined with software such as OpenBugs. The advantage here is that complicated and relatively unique models without dedicated software can be built and estimated in a relatively simple way, using scripts that are both transparent and easily shared with other researchers. The conclusion is that the combination of IRT and generalizability theory is a worthwhile and recommendable methodology for other studies involving variables measured at different levels (students, classrooms), with observations made by different raters at different time points, and using itemized measurement instruments.

Aside from the above-mentioned contributions, the present study also has some shortcomings. One of them is that the relationship between DBDM and DI was not
examined. Based on the DBDM literature, it was assumed that DBDM could result in more data-based DI practices in classrooms and that, if this is the case, student achievement would consequently improve. If DBDM does not result in more data-based DI practices, then DI does not explain (potential) student achievement growth. So, our findings would have contributed more to our understanding of how DBDM influences achievement, if the relationship between DBDM and DI could also have been examined. In addition to this, the Focus intervention was based on the DBDM literature and not on the DI literature. As a result, some effective DI practices unfortunately were not included in the intervention. Ability-grouping effects are, for example, stronger if student grouping is based on mixed data sources like classroom observations, student interviews, and achievement data (Lou et al., 1996), whereas teachers in this study were trained to particularly use achievement data for their group compositions. Differentiated instruction requires more information than students’ results on standardized assessments (National Research Council, 2001). One can think of more qualitative information which teachers collect on a daily basis in the classroom and of the results of diagnostic tests as examples of information sources that can be used for matching instruction with what students need. Furthermore, ability-grouping effects are stronger if group composition is flexible and students do not always stay in the same group (Deunk et al., 2015; Slavin, 1987), whereas teachers in this study were trained to compose ability groups only once every half year in their instructional plan. However, teachers in the intervention were intensively trained at applying other characteristics of effective ability grouping, like composing groups on the basis of specific subject-matter topics and also to vary learning materials and instruction time between ability groups (Deunk et al., 2015; Lou et al., 1996; Slavin, 1987). A second shortcoming of the present study is that we were not able to account for teachers’ colleagues effects. Two (or sometimes even three) teachers sharing classes is quite common in Dutch primary education, due to the high percentage of part-time jobs. Even though teachers who taught mathematics most were asked to participate, and whole school teams were trained, the high percentage of teacher colleagues in both grades still impacted the findings. It is important to take account of this in the interpretation of the study results. A third shortcoming is that teachers’ DI practices observed in the classroom were measured with three items (variation between students’ learning time, instruction activities, and students’ assignments), and more items will be needed to obtain a more validated measure of DI. Another limitation of this study is the validity of the instructional plan checklist. Low and negative correlations between planned DI and DI practices and pretest and posttest scores do not contribute to such validity, so the findings regarding the planning of DI should be interpreted carefully. However, the low correlation between planned DI and DI practices might also be the result of the fact that most participating schools already had finished the Focus intervention. Perhaps the teachers of those schools were, without the support of the trainer, no longer motivated to develop the instructional plan in the way they had learned during the intervention.

Based on the high numbers of classroom observations and raters, in combination with data analysis procedures based on the generalizability theory and the item response theory, our study revealed some valuable insights. Overall, the present study enhances our understanding of DI. The expected variables to be responsible for an achievement effect in a DBDM context were not confirmed by our findings, as no effects
of planned DI and DI practices executed in the classroom were found. However, this study contributed to our knowledge of the effectiveness of ability grouping and points to the importance of further research into why DI and ability grouping seems to be least effective for students in low-ability groups. Based on our findings, we recommend valid measures of DI practices in classrooms, such that not only capture the variation in instruction but also the degree of appropriateness of observed instructional variation for students’ instructional needs.

Disclosure statement

No potential conflict of interest was reported by the authors.

Notes on contributors

Janke M. Faber is a PhD student at the University of Twente. Her research focuses on the evaluation of various manifestations of data-based decision making.

Cees A. W. Glas is professor at the University of Twente. His research focuses on the development and application of statistical models in educational measurement, psychological testing, and large-scale surveys.

Adrie J. Visscher is full professor at the University of Twente and also holds an endowed chair in data-based decision making at the University of Groningen. His research focuses on feedback about the features of teaching activities and feedback about teachers’ impact on student achievement.

ORCID

Janke M. Faber http://orcid.org/0000-0001-8127-6831

References


Appendix 1. Checklist instructional plan

A checklist was developed to measure the quality of DI as planned by teachers. This checklist includes four topics (instruction, learning goals, evaluation, instruction for students with specific learning needs). These topics cover 18 subtopics, which are measured by means of 43 items. Items
were rated by the raters as “present” (1 score), or “not present” (0 score) in the instructional plan. In this Appendix, 18 subtopics of the checklist are presented. More information on this checklist and the translation of the other items can be obtained from the corresponding author.

1. Instruction
   Ability-group composition

   1.1. The composition of the ability groups is in line with the results of the analyses of students’ assessments results, and students’ learning needs (2 items)

Description of instruction

   1.2. The teacher(s) has/have described an instructional approach to address students’ learning needs (3 items)
   1.3. This instructional approach differs between ability groups (3 items)
   1.4. Students’ assignments, learning materials, and/or learning time differ between ability groups (3 items)
   1.5. The teacher(s) has/have specified how assignments and/or learning materials will be used, and for which students (3 items)

2. Learning goals
   2.1. The teacher(s) has/have formulated performance goals for each ability group (a minimum percentage of correct answers for curriculum-based assessments (unstandardized assessments) (1 item)
   2.2. The teacher(s) has/have formulated performance goals for each ability group (a minimum score growth on a standardized assessment) (3 items)

3. Evaluation
   3.1 The teacher(s) has/have formulated how, and when the results of the instructional plan will be evaluated (4 items)
   3.2 The teachers(s) has/have formulated teacher actions in case the learning goals will not be met (1 item)

4. Instruction for students with specific learning needs
   4.1. Teachers’ selection of students with specific learning needs is supported by the results of the analyses of students’ assessment (curriculum-based, and/or standardized assessments), and/or the results of an observation instrument (3 items)
   4.2. Teachers’ description of students’ learning needs is supported by the results of the analyses of students’ assessment results (curriculum-based, and/or standardized assessments) (2 items)
   4.3. Formulated learning goals are in line with the description of students’ learning needs (2 items)
   4.4. Learning goals are formulated in a SMART (specific, measurable, attainable, realistic, timely) way by the teacher(s) (3 items)
   4.5. Summative assessment learning goals are formulated in a SMART way by the teacher(s) (3 items)
   4.6. The teacher(s) has/have given a description of the specific instructional approach (1 item)
   4.7. The teacher(s) has/have given a description of the learning strategies students should learn (1 item)
   4.8. Learning materials are mentioned just as a description is given of how the learning materials will be used by the teacher(s) (1 item)
   4.9. The teacher(s) has/have formulated the organization of instruction for students with specific learning needs (4 items)