Teaching and Teacher Education 64 (2017) 187–198

Contents lists available at ScienceDirect

Teaching and Teacher Education

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

Changes in educators' data literacy during a data-based decision making intervention

Marieke van Geel [a, *], Trynke Keuning [a, b], Adrie Visscher [c], Jean-Paul Fox [a]

[a] Department of Research Methodology, Measurement and Data Analysis, University of Twente, P.O. Box 217, 7500 AE Enschede, The Netherlands
[b] Department Educational Development and Research, Maastricht University, The Netherlands
[c] Department ELAN, University of Twente, P.O. Box 217, 7500 AE Enschede, The Netherlands

ARTICLE INFO

Article history:
Received 12 July 2016
Received in revised form
10 February 2017
Accepted 17 February 2017

Keywords:
Decision making
Program effectiveness
Professional development
Intervention
Data use

ABSTRACT

Data literacy is assumed to be a precondition for the effective implementation of data-based decision making in schools. This study was aimed at investigating changes in 1182 educators' data literacy with regard to student monitoring system data, during a 2-year intervention, which was assessed by using a pretest and posttest.

A multivariate multi-level IRT analysis was conducted. The multivariate approach enabled the identification of differences in initial data literacy and development, based on educators' characteristics. Findings showed significant improvements in educators' data literacy. Furthermore, the 'knowledge gap' between educators with a master's degree versus higher education was closed, just as the gap between teachers and school leaders.

1. Introduction

Although schools are increasingly expected to use data to guide their education, many educators do not feel prepared to use data to inform their practice (Earl & Fullan, 2003; Ikemoto & Marsh, 2007), struggle with the use of data (Huguet, Marsh, & Farrell, 2014), and have shown to lack testing and measurement knowledge required for effective data use (Daniel & King, 1998; Oláh, Lawrence, & Riggan, 2010; Supovitz, 2012). Relatively little attention is dedicated to the preparation of educators in the use of data during their pre-service training (Mandinach & Gummer, 2013a; Mandinach, Gummer, & Muller, 2011; Popham, 2011). Thus, in order to develop their "human capacity to use data", professional development is essential (Mandinach & Gummer, 2013b, p. 21).

In the Netherlands, a comprehensive intervention aimed at implementing data-based decision making (DBDM) was developed and implemented in 101 primary schools. Development of participants' data literacy was stimulated throughout the intervention. This study focused on investigating changes in participants' data literacy as a result of the DBDM-intervention, and at exploring differences in initial scores and the changes in scores, based on educators' characteristics.

2. Theoretical framework

First, the DBDM-intervention will be described shortly. In the sections thereafter, the conceptual framework with regard to 'data literacy for teaching' is discussed, just as the evidence on data literacy development and its effects. An operational definition, taking the context of primary education in the Netherlands into account, is developed for the purpose of this study. The section ends with an overview of the research questions and hypotheses.

* Corresponding author.
E-mail addresses: marieke.vangeel@utwente.nl (M. van Geel), t.keuning@utwente.nl (T. Keuning), a.j.visscher@utwente.nl (A. Visscher), j.p.fox@utwente.nl (J.-P. Fox).

http://dx.doi.org/10.1016/j.tate.2017.02.015
0742-051X/

2.1. The intervention

This study was conducted within the context of a comprehensive professional development intervention: a two-year training course for entire primary school teams, aimed at developing the knowledge and skills for data-based decision making, and implementing and sustaining DBDM in the school organization. A schematic overview of DBDM is depicted in Fig. 1 (van Geel, Keuning, Visscher, & Fox, 2016). DBDM is intended to be implemented as a systematic approach. At class, school and board levels, data are supposed to be analyzed, and these analyses form the basis for setting goals, adapting instruction, adapting the curriculum, evaluating the effectiveness of programs and practices, improving policy, and reallocating time and resources as necessary (Earl & Katz, 2006; Hamilton et al., 2009; Ikemoto & Marsh, 2007; Mandinach et al., 2011). The final step is to implement and execute the chosen strategies. Furthermore, data are also supposed to be used for monitoring and evaluating the effectiveness and outcomes of the implemented actions.

As Mandinach and Gummer (2016) describe, data literacy plays an important role in all steps of the inquiry cycle. Throughout the DBDM-intervention, participants' data literacy was stimulated. By means of workshops on tests, scores, and analyses, participants learned the value of different sources of data and how to interpret these. They furthermore learned how to use the student monitoring system (SMS) and interpret SMS output. Student performance from the SMS was compared to other sources of data, such as curriculum based tests, classroom observations, and diagnostic conversations. Subsequently, participants drew conclusions for improving education and developed (instructional) plans based on their analyses. These plans were executed in practice, and evaluated by means of new data analyses. Participants were required to analyze the performance data of their own students five times during the two intervention years by following a data analysis protocol, and they received individualized feedback on the results of their analyses and the plans they developed. Trainers also devoted time during project meetings to discuss common interpretation mistakes with the entire school team. Twice per school year, schoolwide student performance analyses and evaluations of goals and plans were discussed in a team meeting.

2.2. The data literacy concept

There is wide-spread agreement about the importance and relevance of educators being knowledgeable about testing, assessment and data, and being able to use data correctly. Mandinach and Gummer (2013b) noticed that the terms 'data literacy' and 'assessment literacy' are often used interchangeably. People often seem to think of only assessment data, when talking about data in general. However, data use does not only concern assessment results, but should involve a wide range of data (Mandinach & Gummer, 2013b).

Assessment literacy is often defined in a statistical or technical manner. In their evaluation of the effects of an instructional module to enhance school personnel's assessment literacy, Zwick et al. (2008) defined it as "understanding of the psychometric and statistical principles, needed for the correct interpretation of standardized test scores" (p.15). Interpreting test scores is a vital component of assessment literacy (Sklar & Zwick, 2009), but Popham (2011) took a broader perspective, which includes the understanding of assessment concepts and procedures that influence educational decisions. This is also reflected in the description used by Koh (2011), in which the emphasis lies more on teachers being competent at developing and using assessment and scoring rubrics, and to master evaluative skills to judge student performance.

The concept of data literacy takes a broader perspective, and comprises an array of knowledge and skills that are assumed to be important for the effective use of data in education. For example, Mandinach, Honey and Light (2006) stated that educators need to be able to transform raw data into actionable knowledge, and therefore that skills such as collecting and organizing data, analyzing and summarizing data, and synthesizing and prioritizing data are required. Mandinach (2012) expanded on this description of data literacy by considering the knowledge and skills required for the interpretation and use of data, and referred to this as 'pedagogical data literacy'. This definition includes the transformation of numbers, statistics and analysis outcomes into instructional strategies that meet the students' needs. Earl and Fullan (2003) stressed that the "process of human interpretation and creating meaning has to happen to change data into information and ultimately into workable knowledge" (p.389).

Although there is no consensus among experts, the majority of participants at a convening of experts organized by Mandinach and Gummer (2013b) regarded assessment literacy as a component of data literacy. The common conflation of data literacy and assessment literacy, and the lack of a common, operational definition led to the development of a conceptual framework on data literacy for teaching by Gummer and Mandinach (2015). They argued that data literacy is closely intertwined with other broad domains of teaching, such as disciplinary knowledge, pedagogical content knowledge, and understanding about student development. In their conceptual framework, the domain of data use for teaching is unpacked and presented as parts of the different steps in the inquiry cycle (Gummer & Mandinach, 2015; Mandinach & Gummer, 2016). At each step in this cycle, from identifying problems to using data, transforming data into information, transforming information into decisions and evaluating outcomes, teachers require specific knowledge and skills to make sense of the data they are using. This knowledge and skills together form the domain of data

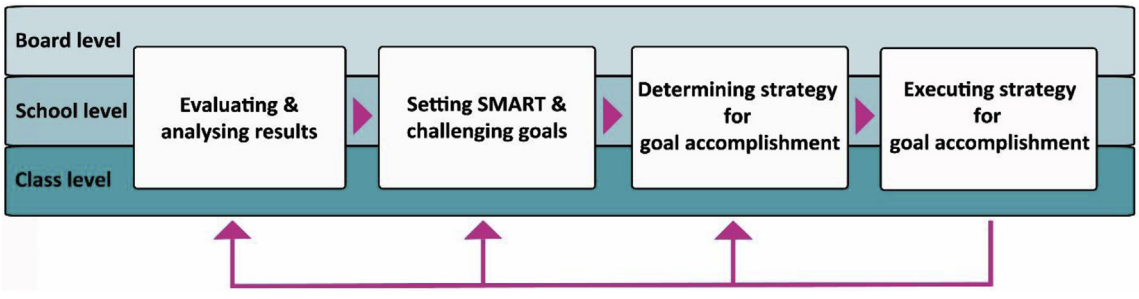

Fig. 1. Schematic overview of DBDM.

literacy for teaching.

2.3. Data literacy evidence

Although the systematic use of data is regarded as an important characteristic of effective teachers, evidence on both the development of data literacy as well as on the contribution of data literacy to student achievement is scarce. Several studies have concluded that teachers' fundamental testing and measurement knowledge is insufficient (Daniel & King, 1998; Oláh et al., 2010; Supovitz, 2012), and that relatively little attention during their pre-service training is dedicated to the preparation of educators for the use of data (Bron, van Geel, & Visscher, 2013; Mandinach et al., 2011; Mandinach & Gummer, 2013a; Popham, 2011).

Empirical studies on developing data literacy points towards a similar conclusion. For example, Zwick et al. (2008) evaluated the effect of an instructional module about psychometric and statistical principles, and found that these modules were effective for students in teacher education preparation programs, but not for school personnel. The scores of in-service teachers, with an average experience of 11 years, were comparable to post-intervention scores achieved by pre-service teachers. Apparently, these teachers developed this knowledge during their work. On the other hand, in the study by Gotch and French (2013) no correlation was found between teaching experience and assessment literacy or assessment self-efficacy. An empirical study by Koh (2011) revealed that teachers who received ongoing, sustained professional development over the course of two school years instead of one-shot workshops, showed significant improvements in the use of authentic assessment, quality of assessment, and student engagement in learning. The importance of professional development to build human capacity for data use is widely acknowledged (Mandinach & Gummer, 2013b, p. 21). Therefore, the professional development intervention as described above was developed at the University of Twente.

2.4. Defining data literacy for this study

As Mandinach and Gummer (2013b) noted, 'data literacy' cannot be defined without taking the landscape of data use into account. For the purpose of this study, we combined insights from general data literacy and assessment literature with the context characteristics of data use in the Netherlands, as a basis for our operationalization of 'data literacy'.

An important aspect in the context of primary education in the Netherlands is the student monitoring system. Almost all primary schools (94 percent, according to the Inspectorate of Education (2016)) use this coherent set of tests for the longitudinal assessment of students' achievement throughout all grades of primary education, which was developed by the Central Institute of Test Development (Kamphuis & Moelands, 2000). These tests, which are usually taken twice a year (in January and in July), are available for all core subjects (mathematics, reading, spelling and vocabulary). The test results are converted into an ability scale for each subject so that student progress can be monitored over grades and school years (Kamphuis & Moelands, 2000). Although the results of some of these tests are used for accountability purposes by the Inspectorate, the tests are clearly designed for monitoring student achievement progress and analyzing patterns in achievement across students and grades. The tests are therefore generally not perceived as 'high-stakes' tests (Kamphuis & Moelands, 2000).

To process the collected data, tools are required to help educators organize, analyze, interpret and report data in meaningful ways (Bernhardt, 2005; Mandinach, 2012). Student monitoring system software is an example of such a technical tool. As mentioned previously, the tests in the student monitoring system enable educators to monitor students' progress throughout their entire school career (Kamphuis & Moelands, 2000; Verhaeghe, Schildkamp, Luyten, & Valcke, 2015). After taking a test, teachers can store the student performance data in the SMS-software. Graphs, tables and growth models representing various aspects of student performance can then immediately be retrieved from the system, and with the SMS-software it also is possible to compare the scores of students and grades with national percentile scores.

Although there is a clear distinction between the student monitoring system (a coherent set of tests) and the SMS-software to analyze the results, in practice the term 'student monitoring system' is used for both of these tools. In the remainder of this paper, 'student monitoring system' and the abbreviation SMS are used for the set of tests as well as the software used to analyze the results.

In the working definition by Mandinach and Gummer (2016), data literacy for teaching entails a range of skills, from being able to interpret student achievement data correctly to using this data to inform practice. In the context of the current study the focus is on a subset of skills from this framework, related to using and interpreting assessment data from the student monitoring system. We focused on teachers' knowledge about what kinds of analyses can and cannot be done with the system, the correct interpretation of graphs, tables, and other data representations, and relating scores to benchmark data. These skills can be found in the framework by Mandinach and Gummer (2016) under the components 'Use Data' (e.g. subcomponents 'understand data properties'; 'understand how to access data'; 'understand how to analyze data'), 'Transform Data into Information' (e.g. element 'understand data displays and representations'), and 'Transform Information into Decision' (e.g. elements 'monitor student performance'; 'diagnosis of students' needs').

2.5. Research question and hypotheses

Since data literacy development of participants was one of the explicit aims of the intervention, the research question was: *can data literacy of educators be improved, by means of an intensive DBDM intervention?*

It was expected that their data literacy levels would improve during the intervention (*hypothesis 1*). Before DBDM was introduced, often only school leaders and academic coaches used data from a SMS to analyze student achievement. Staman, Visscher and Luyten (2014) found that initial scores for school leaders and internal coaches were significantly higher than for teachers, a finding we expect to replicate (*hypothesis 2*). Institutes for teacher training in the Netherlands are institutes for higher education (universities of applied science), with graduate level comparable to a bachelor's degree, although the government stimulates educators to attain a master's degree as well. It was expected that data literacy of educators with a master's degree would initially be higher, since they would be more familiar with statistical concepts and graphical data representations as compared to educators with higher education – only limited attention is payed to statistics in higher education curricula (*hypothesis 3*). Furthermore, it was expected that educators – especially prior to the intervention – would have benefited from each other's expertise. As such, the general level of data literacy would be higher in schools in which teams consisted of more highly educated staff. The proportion of team members that completed a master-level education was therefore expected to be positively associated with data literacy (*hypothesis 4*). The intervention was aimed at reaching the same level of data literacy for all participants, it was therefore expected that the 'gap' in data literacy ability, based on educational level and function (*hypothesis 5 and 6*)

would decrease. Furthermore, scores were controlled for school background variables. Differences might be attributed to cohort effects, since half of the schools started the intervention a school year later. We also controlled for student monitoring system.

3. Method

3.1. Data collection

In order to determine participants' data literacy levels, a SMS data literacy test was administered at both the beginning and end of the intervention. Since three SMS software systems (C, E, and P) are commonly used by schools in the Netherlands, three test versions were developed. The tests consisted of a set of general items for all participants, and system specific questions for each of the SMS software systems. Note that all three systems are used to represent the scores on the same SMS tests, and that schools are free to choose the program they prefer. The SMS data literacy tests were developed in the summer of 2010 to be used in the first cohort of the project (not included in this study). In the tests, data literacy entails "knowing what kinds of analyses can and cannot be made with the system, and interpreting graphs, tables, and other data representations correctly". The tests therefore consisted of two parts: a) knowing which analyses can be made with the system (the 13 general data literacy items about possible analyses), and b) the correct interpretation of results from the system (the two general items about the interpretation of benchmarks, and the items about the interpretation of software specific representations). Questions were asked about the interpretation of outcomes from the analyses at the school level, the classroom level, and the individual student-level. Data literacy comprises many subskills, but was regarded as one latent variable in the analysis of this test.

The 'knowing which analyses are possible' items were based on general SMS software possibilities and common (mis)conceptions about possibilities and interpretations. The two items about the general interpretation of benchmarks were meant to measure the interpretation and use of the letters indicating student performance levels, related to the national average.[1] For the software specific interpretation items, data representations were used that were expected to be common in DBDM practice, such as cross sections, trend analyses, and overviews of achievement growth. In Table 1, the number of general items and system specific items per SMS test version are depicted.

The general items were shared across test versions. Some items were about the general possibilities of the SMS system software, for which the answer sometimes was the same across systems (for example: "With the SMS, we can compare the scores of our students with students in other countries, is this true or false?", which is false for all systems), and sometimes the correct answer differed across systems (for example: "With the SMS, we can monitor student scores for different subareas of the test, is this true or false?", which is true for systems C and E, but false for system P). Two items were removed from the test prior to the analyses in this study, because the correct answers changed over time due to software updates. This left thirteen general data literacy items about possibilities for the analysis. Out of these thirteen items, two had different correct answers for different SMS's and were therefore treated as system specific items.

The two other general items were not related to the software system but about the general interpretation of performance levels (for example: "Margot scored a level B on the previous test and now scored a level C. What happened to her score, relative to the national average?"). These items are labeled as 'interpreting benchmarks' in Table 1.

The items about analyses and representations in the specific systems related to software possibilities were about the interpretation of graphical representations of test results in the software system (for example: "Based on the trend graph above, what can you say about the scores of grade 3 in school year 2011-'12, compared to their scores in the previous school year?"). Because the specific software possibilities differed across systems, the number of items per system differed as well. The items about the interpretation of analyses in the different systems were kept as similar as possible, using representations of roughly the same information and asking the same types of questions about these analyses.

Since the graphical data representations in system P changed over time, two versions of this data literacy test were developed. Test version P-pre was used as a pre-test, version P-post was used as a post-test. Three out of ten system specific items for system P were identical in both versions.

The thirteen general items about possibilities with the systems were statements to which the respondents could respond by choosing from true, false or I don't know. The two general items about interpretation and benchmarks, and the items about analyses in the specific systems were multiple choice questions, with three or four answer choices, and the option "I don't know". Responses were scored dichotomously: 1 for a correct response and 0 for an incorrect or missing response. The response "I don't know" also was scored as zero.

Aside from the data literacy test, participants' background characteristics such as gender, age, and educational level, were collected by means of a questionnaire.

3.2. Sample

For this study, two cohorts of schools were used. The first cohort, of 53 schools, was exposed to the intervention in the school years 2011-12-13, the second cohort, of 48 schools, in 2012-13-14. The data literacy pretest was conducted at the start of the intervention, in August 2011 (for cohort 1) and August 2012 (for cohort 2). Posttests were conducted at the end of the intervention – in July 2013 (cohort 1) and July 2014 (cohort 2). Out of the total number of 1883 unique respondents, 1204 (64%) completed both the pre-test and the post-test, 445 respondents (24%) only completed the pre-test, and 234 (12%) only completed the posttest. For the purpose of this study, only respondents who completed both pre-test and post-test were included. Furthermore, only schools with five or more respondents were included. This lead to a total of 1182 respondents from 83 schools. An overview of the number of respondents, and schools per system and measurement occasion can be found in Table 2.

In Table 3, respondents' characteristics are presented. The majority of respondents were teachers and female, approximately two-third of the respondents completed higher education, which is the common educational level for primary school teachers in the Netherlands. Since only 2.4% of the respondents were older than 60 years at the beginning of the intervention, the highest age category was set to include participants aged 51 years and older.

3.3. Data analysis

3.3.1. Conceptual model

It was assumed that the data literacy test can be used to measure

[1] The letters indicating levels are based on the distribution of student achievement. A = top 25%; B = 25% above average; C = 25% below average; D = 15% far below average; E = lowest scoring 10%. The national average lies between level B and C.

Table 1
Overview of the number of items per category per data literacy test version.

Test version	General Data Literacy			Analyses in system C	Analyses in system E	Analyses in system P	
	Possibilities	Interpreting benchmarks	System specific				
SMS C	11	2	2	15	0	0	
SMS E			2	0	11	0	
SMS P-pre			2	0	0	3	7
SMS P-post				0	0		7

Table 2
Overview of final number of respondents (and schools) per system and measurement occasion.

	Pre-test			
	SMS C	SMS E	SMS P	Total
Cohort1	342 (24)	56 (3)	222 (18)	620 (45)
Cohort2	165 (13)	124 (9)	273 (16)	5621 (38)
Total	507 (37)	180 (12)	495 (34)	1182 (83)

Table 3
Overview of respondents' characteristics (final sample).

	Respondents' characteristics (N = 1182)
Gender	
Male	14.1%
Female	85.9%
Education (highest level)	
Master's degree	30.2%
Higher Education	60.3%
Education	8.7%
Function (highest)	
School leader	8.9%
Internal coach	7.9%
Teacher	78.5%
Other	3.6%
Age (at pretest)	
≤ 30 years	20.7%
31–40 years	24.5%
41–50 years	22.1%
≥ 51 years	32.8%

the unidimensional, latent ability 'SMS data literacy'. However, because the tests consisted of three parts, of which two were administered to all participants and one was SMS software specific, raw scores were neither suitable for comparing scores of users across different systems nor for comparing achievement over time. For example, it is possible that items that were specific for system C were easier than the items specific for system P, e.g. because the graphical representations in the former system were easier to interpret than those in the latter. Furthermore, educators were nested within schools, requiring a multilevel analysis.

Fig. 2 depicts the conceptual model applied in this study. At both the individual level as well as at the school level, pre-test and post-test score correlated. Covariates at the individual level were gender, age, educational level and function. At the school level, the effects of cohort, proportion of master's, and the student monitoring system used by the school were included.

3.3.2. Multivariate multi-level pre-post IRT model

In order to enable comparisons within and between respondents, take the nested structure into account and to establish different covariate effects for pre-test and post-test, a multi-level IRT analysis, with a multivariate approach of the pre-post design, was conducted.

As depicted in Fig. 2, the latent variable 'data literacy' was measured at the educator level, with educators nested within schools. This data structure required a multilevel analysis, since item responses were clustered within educators, who in turn were clustered within schools. Data literacy at the time of the pre-test and post-test was measured using a two-parameter item response model, while accounting for the multilevel design, representing educators nested in schools, and item test differences across measurement occasions.

For the pre-test, the difficulty and discrimination parameters for item k ($k = 1, \ldots, k^{pre}$) are denoted as b_k^{pre} and a_k^{pre}, respectively. The data literacy of educator i in school j at the pre-test is denoted as θ_{ij}^{pre}. The superscript *post* is used to refer to post-test characteristics. The pre-test and post-test item response models were used to measure the educator's data-literacy ability. The probability of a correct response to item k on the pre-test and the post-test, by educator i of school j is given by

$$P\left(Y_{ijk}^{pre} = 1 \middle| \theta_{ij}^{pre}, a_k^{pre}, b_k^{pre}\right) = \Phi\left(a_k^{pre}\theta_{ij}^{pre} - b_k^{pre}\right),$$
$$P\left(Y_{ijk}^{post} = 1 \middle| \theta_{ij}^{post}, a_k^{post}, b_k^{post}\right) = \Phi\left(a_k^{post}\theta_{ij}^{post} - b_k^{post}\right),$$

respectively. The $\Phi(.)$ denotes the cumulative normal distribution function, and for each item response model the success probability depends on the occasion-specific item parameters and the educator's data-literacy ability.

To ensure that the pre-test and post-test data-literacy latent variables were measured on the same scale, anchor items were selected for which parameters were fixed across measurement occasions. The anchor items were used to link the pre-test and post-test scales. Furthermore, educators and schools were measured on both occasions, which led to correlated measurements at the level of educators and schools. Therefore, as depicted in Fig. 2, pre- and post-test measurements were assumed to be correlated at the school level as well as the educator level. A multivariate multilevel modeling approach was defined to control for these correlations. At the level of educators, a multivariate distribution was assumed for the latent variables for the pre- and post-test measurement, which is given by

$$\theta_{ij}^{pre} = \beta_{0j}^{pre} + X_{ij}^{pre}\beta_1^{pre} + e_{ij}^{pre}$$
$$\theta_{ij}^{post} = \beta_{0j}^{post} + X_{ij}^{post}\beta_1^{post} + e_{ij}^{post}$$

where the educator and school-level explanatory variables are given by $\mathbf{X}^{pre}$ and $\mathbf{X}^{post}$ for both occasions. The error terms were assumed to be multivariate normally distributed to model the correlation between the educator's data-literacy ability at the pre- and post-measurement occasion. That is;

$$\begin{pmatrix} e_{ij}^{pre} \\ e_{ij}^{post} \end{pmatrix} \sim N\left(\begin{pmatrix} 0 \\ 0 \end{pmatrix}, \begin{pmatrix} \tau_{pre}^2 & \tau_\theta \\ \tau_\theta & \tau_{post}^2 \end{pmatrix}\right),$$

where the occasion-specific parameters τ_{pre}^2 and τ_{post}^2 represent the error variances. The parameter τ_θ represents the covariance

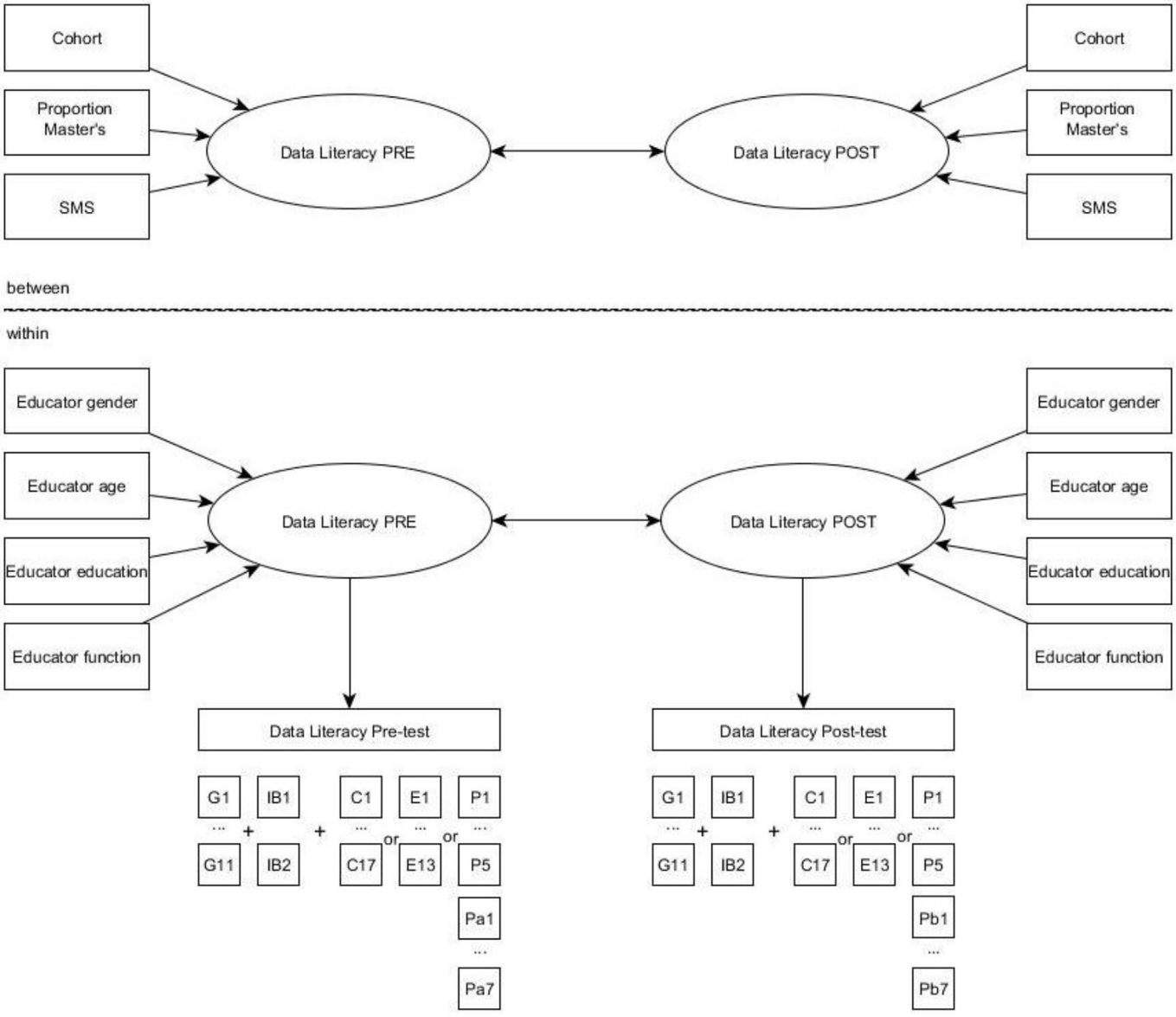

Fig. 2. Conceptual model.

between educator's measurements, while controlling for differences by the mean terms including the explanatory variables.

The random school components $(\beta_{0j}^{pre}, \beta_{0j}^{post})$ describe the average performance of the school's educators at both occasions, while controlling for differences between educators and controlling for possible differences in effects of the occasion-specific explanatory variables. The random school components were also assumed to be multivariate normally distributed to address the correlation between the pre-test and post-test measurements of the same school. This multivariate level-2 part is given by

$$\begin{pmatrix} \beta_{0j}^{pre} \\ \beta_{0j}^{post} \end{pmatrix} \sim N\left(\begin{pmatrix} \gamma_0^{pre} \\ \gamma_0^{post} \end{pmatrix}, \begin{pmatrix} \sigma_{pre}^2 & \sigma_\beta \\ \sigma_\beta & \sigma_{post}^2 \end{pmatrix} \right),$$

where γ_0^{pre} and γ_0^{post} are the expected population-average scores on the pre-test and post-test, respectively. The covariance matrix represents the variability in school's data literacy on the pre-test and post-test, and σ_β represents the covariance between pre and post-test scores at the school level.

The pre and post-test measurements are defined on the same scale, since anchor items are used to link the scales. The scale is identified by fixing the mean to zero and variance to one for the pre-test. This is a common identifying restriction in multiple-group IRT modeling (Azevedo, Andrade, & Fox, 2012; Bock & Zimowski, 1997; Reise, Widaman, & Pugh, 1993). Therefore, $\gamma_0^{pre} = 0$ and the total variance of the pre-test measurements is restricted to one, where the total variance is represented by the variance components at the different levels. The restriction on the total latent variance for the pretest is included in the estimation method. That is, the pre-test measurements are rescaled to have a mean of zero and a variance of one during the estimation of the model parameters. Since the parameter γ_0^{pre} is fixed to zero, the parameter γ_0^{post} represents the average pre-post effect, representing the average score differences between the pre- and post-measurements given the explanatory variables.

The multivariate two-level model for the latent pre- and post-test measurement can be stated as a multivariate regression

model with error components at different levels; that is,

$$\left(\theta_{ij}^{pre}, \theta_{ij}^{post}\right) = \left(\mathbf{X}_{ij}^{pre}\boldsymbol{\beta}^{pre}, \mathbf{X}_{ij}^{post}\boldsymbol{\beta}^{post}\right) + \left(u_{0j}^{pre}, u_{0j}^{post}\right) + \left(e_{ij}^{pre}, e_{ij}^{post}\right),$$

where the two error components at the educator and school level are multivariate normally distributed to model the correlation between educator's measurements and school measurements. The school and educator measurements are on a common scale due to the anchor items, and due to the correlation between measurements.

Schools and educators within each school were tested during the pre-test and post-test. Therefore, the covariance between those measurements is complex since educators are nested in schools. The covariance between the pretest and posttest measurements of school j is given by

$$Var\left(\boldsymbol{\theta}_j^{pre}, \boldsymbol{\theta}_j^{post}\right) = \begin{bmatrix} \boldsymbol{\Sigma}_{11} & \boldsymbol{\Sigma}_{12} \\ \boldsymbol{\Sigma}_{21} & \boldsymbol{\Sigma}_{22} \end{bmatrix} = \begin{bmatrix} \tau_{pre}^2\mathbf{I}_{nj} + \sigma_{pre}^2\mathbf{J}_{nj} & \tau_\theta\mathbf{I}_{nj} + \sigma_\beta\mathbf{J}_n \\ \tau_\theta\mathbf{I}_{nj} + \sigma_\beta\mathbf{J}_{nj} & \tau_{post}^2\mathbf{I}_{nj} + \sigma_{post}^2\mathbf{J}_{nj} \end{bmatrix}$$

where $\mathbf{I}_{n_j}$ is the identity matrix of dimension n_j and $\mathbf{J}_{n_j}$ a matrix of ones of dimension n_j. The covariance matrices $\boldsymbol{\Sigma}_{11}$ and $\boldsymbol{\Sigma}_{22}$ define the correlation between pre-test and post-test measurements due to the nesting of educators in schools, respectively. The covariance matrix $\boldsymbol{\Sigma}_{12}$, which represents the covariance between the posttest and pretest measurements of school j, has on the diagonal τ_θ to model the correlation between each educator's pretest and posttest measurement. The non-diagonal terms describe the covariance between measurements of school j at the pre- and post-test.

In this multivariate multilevel modeling approach, pre- and post-test measurements are jointly modeled at each (hierarchical) level of the model, which means that educator and school pre-test and post-test measurements are each jointly modeled. Here, the repeated measurements, made at both modeling levels, are also clustered in a cross-sectional way at each measurement occasion. Both pre-post models extend the multilevel item response model of Fox and Glas (2001) and Fox (2010).

Multivariate analysis of pre- and post-test scores has the advantage that covariates (such as educational level, gender and age) can be added independently at both measurement occasions. The effect therefore can also differ across occasions, and the effects of these variables at both pre-test and post-test can be compared. Furthermore, in multivariate analysis the measurement errors of both measurement occasions are taken into account. Differential measurement errors are allowed at the level of schools and educators. It is also possible to include time-invariant and time-variant explanatory variables, which can differ in their effects on the pre and post measurement.

In the traditional change score method (e.g. Allison, 1990), the difference score $\beta_{0j}^{post} - \beta_{0j}^{pre}$ or $\theta_{ij}^{post} - \theta_{ij}^{pre}$ would be considered as a dependent variable, since the change score method is not developed for clustered outcomes. The linear regression of the change scores on the predictors ignores the differential measurement error of educator and/or school scores. This will lead to biased estimates of pre-post effects. Furthermore, when effects of explanatory variables are not invariant from the pre to posttest, they need to be included to account for differences between the two occasions. In the change score method, it is only possible to address effects of pre-posttest differences in explanatory variables, but not of the actual effects at each occasion. This complicates the interpretation of the effects of explanatory variables. Time-invariant explanatory variables will drop out of the equation, although their effects might be different for the pre and posttest.

In the so-called regressor variable approach, the pre measurement is used as a control variable, which implies that β_{0j}^{post} is regressed on β_{0j}^{pre} (or of θ_{ij}^{post} on θ_{ij}^{pre}) and explanatory variables, to make inferences about changes in school performances. Again, the regressor variable approach has not been developed for clustered outcomes. It also follows that the measurement error associated with the pre measurement can produce biased estimates of the pre-post effect. Furthermore, in correspondence to the changes score method, the inclusion of explanatory variables is restricted to changes in the variables measured at both occasions. The regressor variable approach is implied by the multivariate model.

When considering the conditional distribution of θ_{ij}^{post} given θ_j^{pre}, it follows that the conditionally expected measurement for educator i in school j at the post test given the pre-test measurements of all educators of school j is given by

$$E\left(\theta_{ij}^{post}\middle|\boldsymbol{\theta}_j^{pre}\right) = \mathbf{X}_{ij}^{post}\beta^{post} + \boldsymbol{\Sigma}_{12}\boldsymbol{\Sigma}_{11}^{-1}\left(\boldsymbol{\theta}_j^{pre} - \mathbf{X}_{ij}^{pre}\beta^{pre}\right), \quad (0.1)$$

where the covariance between pre and post measurement is defined by the covariance between the educator's measurements but also between the school's measurements. So, beside the covariance between educator's measurements, the covariance between the school's measurements contribute to the covariance between an educator's post-test measurement and the pre-test measurements of school j. School j is measured at the pre-test and the post-test, which lead to correlated measurements of school j. Subsequently, each educator's post-test measurement is related to all pre-test measurements of school j.

3.3.2.1. Statistical inferences from the pre-post model. Following the procedure as described, Bayesian inferences can be made by computing characteristics of the posterior distribution of the model parameters. The estimated posterior means are used as estimates of the model parameters. Bayesian 95% confidence intervals (i.e., highest posterior density (HPD) intervals) can be calculated, where each point inside the interval has a higher posterior density value than excluded points (e.g. Fox, 2010, p. 59). The 95% HPD intervals can be used to evaluate whether estimated parameters are significantly different from zero.[2] This procedure is used to evaluate whether effects of predictor variables are significantly different from zero. Explanatory variables without a significant effect on both pre- and posttest factor scores, can be excluded from the model. Furthermore, the estimated characteristics of the posterior distribution of the difference between pre-test and post-test parameters can be used to evaluate whether predictor effects differ across occasions. It follows that this procedure is specifically useful in multivariate multilevel modeling, where an interest is in the differential effects of predictors across measurement occasions.

The empty model can be used to investigate the amount of level-1 variance between educators and level-2 variance between schools. By considering the cross-sectional clustering at the pre and post-test occasion, the intra-class correlation coefficient at the pre and post measurement can be defined as

$$\rho_{pre} = \frac{\sigma_{pre}^2}{\sigma_{pre}^2 + \tau_{pre}^2}, \text{ and } \rho_{post} = \frac{\sigma_{post}^2}{\sigma_{post}^2 + \tau_{post}^2}.$$

[2] The (Bayesian) HPD interval provides information about the most likely parameter values given the posterior information. The true parameter value is with 95% probability likely to be different from zero when zero is not included in the parameter's 95% HPD interval. This will be referred to as a significant effect.

Table 4
Results simulation study for the parameter recovery of the pre post multivariate multilevel model.

	True values		Estimated values (over 50 replications)			
	pre	post	pre		post	
			Mean	SD	Mean	SD
Fixed effects						
Intercept	0.000	0.047	0.000	0.001	0.040	0.036
Covariate 1	1.049	0.888	1.031	0.080	0.856	0.085
Random effects						
Lvl-1 variance	1.0	0.8	1.016	0.057	0.796	0.098
Lvl-2 variance	0.5	0.3	0.511	0.093	0.314	0.068
Covariance lvl-1	0.30		0.241	0.040		
Correlation lvl-1	0.34		0.268	0.037		
Covariance lvl-2	0.15		0.178	0.049		
Correlation lvl-2	0.39		0.450	0.094		

They represent the proportional amount of variance explained by the clustering of educators in schools. Following Snijders and Bosker (1999, pp. 102–103), the proportional reduction in variance at each level can be computed to quantify the relevance of explanatory variables included in the model, at the level of educators or schools.

3.3.3. Simulation study

A Markov Chain Monte Carlo (MCMC) algorithm was developed to estimate all parameters, which was also used to obtain sampled values of functions of parameters. This was used to compute the intra-class correlations and the proportional reductions in level-1 and level-2 variances. The MCMC algorithm is an extension of the MCMC methods for multilevel IRT models (e.g., Fox & Glas, 2001; Fox 2010; Stone & Zhu, 2015) and the pre-post multivariate multilevel IRT model.

A simulation study was carried out to investigate the parameter recovery performance of the developed MCMC algorithm. Therefore, a total of 50 data sets were simulated according to the model with true parameter values as given in Table 4. Each simulated data set consisted of 1000 educators, nested in 100 schools, and educators' dichotomous item responses were generated for 10 items on the pre and posttest. For each data set, the MCMC algorithm was ran for 5000 iterations, a burn-in period of 1000 iterations was used, and the remaining sampled values from the posterior distributions were used to estimate the model parameters and standard deviations. The reported estimates in Table 4 are the pooled values, where the average is taken over the estimates corresponding to the 50 generated data sets.

It can be seen from Table 4 that the true parameter values were well recovered. It was concluded that the developed MCMC algorithm produced correct estimates for the parameters of the pre-post multivariate multilevel IRT model.

3.3.4. Missing data

Missing answers were scored as incorrect. Furthermore, a missing data matrix was defined, indicating missing by design for the system E and P specific items for respondents who worked at schools with system C, and so on for users of system E and P. Missing data with regards to covariates were replaced with the most common value: female, teacher, higher education and $\geq$51 years old.

3.3.5. Inclusion of items

The first step was to determine whether discrimination and difficulty parameters were acceptable for all items. Based on an empty model with all general items set as invariant, items were removed when discrimination parameters were below 0.30 at all moments the item was administered (three items in total) or when the difficulty was below −3 or higher than 3 (one item). In Table 5, the final number of items per category for each test version is presented.

3.3.6. Selection of anchor items

After determining the final set of items, the items that would be set as invariant across measurements were selected. It was the aim to select the five general items with the highest factor loadings as anchor items. The stability of the intended anchor items was tested by comparing item parameter estimates across different combinations of anchor items, selected from all general items. There was some variability detected in item parameter estimates over different anchor settings, which is partly caused by the sample size and sampling error. The mean estimated posterior standard deviation of the difference in item parameter estimates was approximately 0.10 and 0.11 for the discrimination and difficulty parameter, respectively. So, some item parameter differences were expected due to sampling error. The maximum difference in discriminations was approximately 0.2. Under one anchor specification, items with a low difficulty value showed more variation in estimates, with a maximum of 0.4. This was caused by the fact that in this setting all specified anchors had low discrimination values, which led to a larger difference in estimated pretest and posttest item difficulties of the non-anchor items, compared to differences obtained with anchors with higher discrimination values. Subsequently, this also led to larger differences in item difficulty estimates when comparing the estimates with those obtained under another anchor setting. As a result, the five items were considered appropriate anchor items.

An extensive description of item parameter estimates and estimates of standard errors of the construct estimates can be found as Supplementary material.

4. Results

4.1. Empty model

The first step was to identify the empty model, in order to explain variance at the school and educator levels at the time of the

Table 5
Overview of the final number of items per category per data literacy test version.

Test version	General Data Literacy			Analyses in system C	Analyses in system E	Analyses in system P	
	Possibilities	Interpreting benchmarks	System specific				
SMS C	10	2	1	15	0	0	
SMS E			1	0	11	0	
SMS P-pre			2	0	0	3	7
SMS P-post				0	0		6

Table 6
Empty model.

	Pre-test				Post-test			
	Estimate	SD	HPD min	HPD max	Estimate	SD	HPD min	HPD max
Fixed effects								
Intercept	0.002	0.054	[-0.101,	0.106]	1.362	0.111	[1.142,	1.580]
Random effects								
Level2 variance (schools)	0.148	0.034	[0.086,	0.214]	0.177	0.049	[0.095,	0.274]
Level1 variance (educators)	0.881	0.038	[0.808,	0.958]	1.435	0.147	[1.176,	1.721]
Intra Class Correlation	0.144				0.110			
Level1 covariance pre-post	0.598	0.052	[0.497,	0.702]				
Level1 correlation pre-post	0.533	0.035	[0.466,	0.601]				
Level2 covariance pre-post	0.102	0.031	[0.046,	0.161]				
Level2 correlation pre-post	0.632	0.133	[0.375,	0.887]				

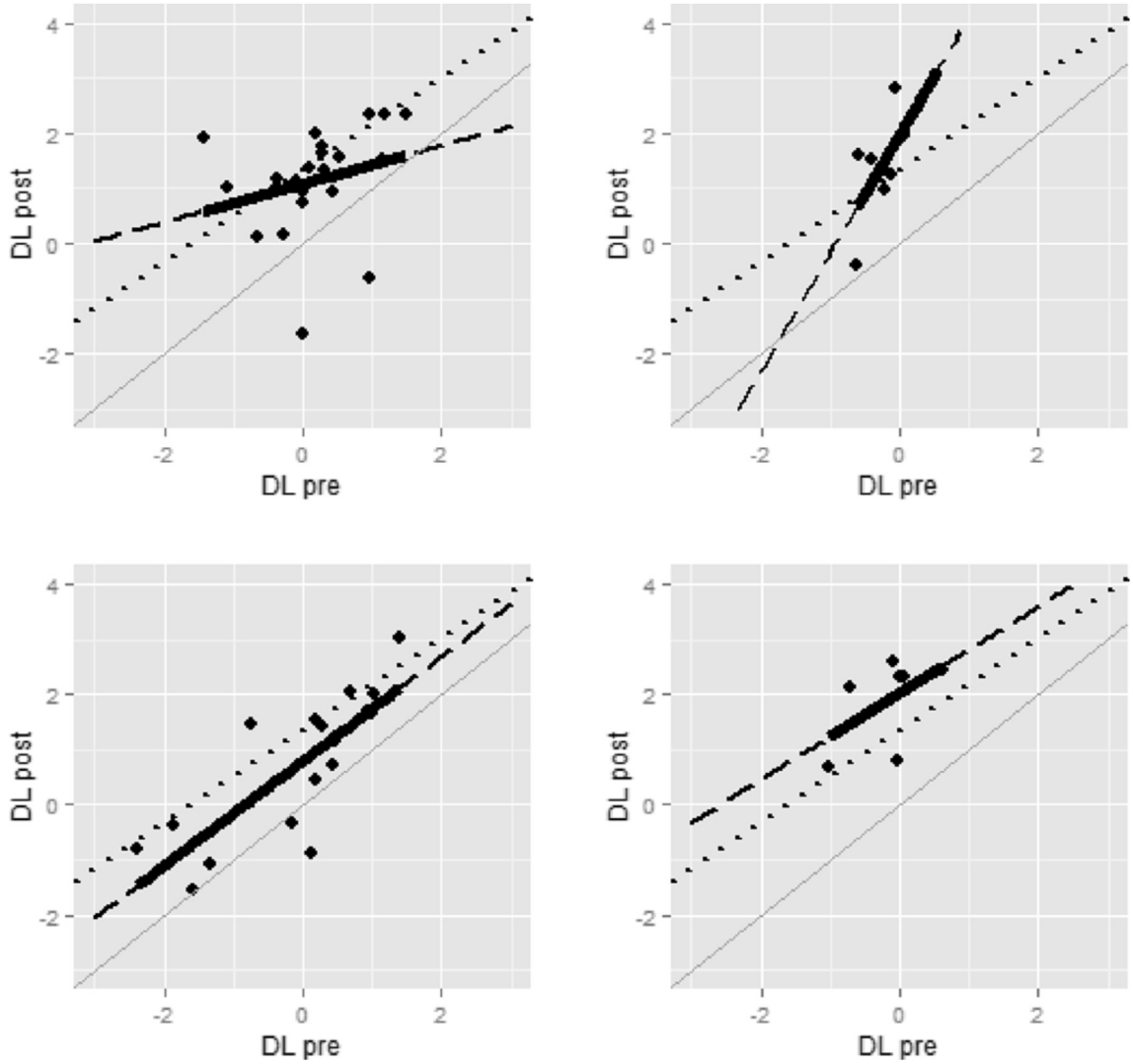

Fig. 3. Pre-test and post-test scores for four different schools. Reference line (grey, straight line) with intercept 0 and slope 1. Dotted line represents population intercept and slope, dashed line represents intercept and slope for selected school, fitted area was printed bold.

pre-test and post-test, and to identify overall differences in pre-test and post-test scores, without taking covariates into account. The results for the empty model are presented in Table 6.

The nested structure of the data, with responses nested in educators, and educators nested in schools, is captured by two random effects per measurement occasion: the random effect at the school level ($\beta_{oj}^{pre}, \beta_{0j}^{post}$) and at the educator level ($\theta_{ij}^{pre}, \theta_{ij}^{post}$). The random effect variances indicate that the largest part of the

Table 7
Final model (Model 5).

	Pre-test				Post-test				Changes in covariate effects			
	Estimate	SD	HPD min	HPD max	Estimate	SD	HPD min	HPD max	Estimate	SD	HPD min	HPD max
Fixed effects (Level2)												
Intercept	−0.077	0.104	−0.275	0.124	1.429	0.185	1.102	1.805				
Covariates (Level1)												
Education − Lower Voc.	−0.933*	0.107	−1.146	−0.740	−0.695*	0.142	−0.975	−0.422	0.238	0.185	−0.126	0.592
Education − Master	0.276*	0.068	0.135	0.399	0.185	0.094	0.000	0.365	−0.091	0.119	−0.327	0.133
Function − Other	−0.035	0.160	−0.348	0.281	−0.369	0.201	−0.776	0.002	−0.334	0.259	−0.840	0.181
Function − Academic coach	0.835*	0.113	0.611	1.056	0.747*	0.159	0.437	1.053	−0.088	0.196	−0.475	0.288
Function − School leader	0.511*	0.108	0.310	0.725	0.162	0.154	−0.139	0.460	−0.349*	0.189	−0.731	−0.002
Gender − female	−0.223*	0.082	−0.384	−0.068	−0.223	0.128	−0.481	0.019	0.000	0.158	−0.301	0.313
Age ≤30	0.161	0.086	−0.009	0.328	0.274*	0.124	0.023	0.514	0.112	0.157	−0.198	0.409
Age 31-40	0.147	0.081	−0.014	0.299	0.129	0.119	−0.117	0.360	−0.019	0.148	−0.299	0.266
Age ≥51	−0.237*	0.076	−0.388	−0.084	−0.472*	0.107	−0.691	−0.270	−0.236	0.134	−0.495	0.024
Covariates (Level2)												
Cohort (FII = ref)	0.328*	0.081	0.167	0.483	0.296*	0.109	0.085	0.510	−0.032	0.123	−0.269	0.210
Random effects												
Level2 variance (schools)	0.094	0.022	0.055	0.135	0.133	0.038	0.069	0.213				
Level1 variance (educators)	0.650	0.030	0.594	0.711	1.201	0.127	0.946	1.432				
Intra Class Correlation	0.126				0.100							
Level1 covariance pre-post	0.393	0.039	0.318	0.471								
Level1 correlation pre-post	0.446	0.036	0.369	0.511								
Level1 pre-post effect	0.606	0.065	0.482	0.732								
Level2 covariance pre-post	0.053	0.021	0.014	0.096								
Level2 correlation pre-post	0.474	0.153	0.161	0.747								
Level2 pre-post effect	0.576	0.235	0.175	1.070								

Note: * indicates that zero is not included in the 95% highest posterior density interval.

variance is explained at the individual level. The level-1 random effect identifies the clustering of responses within subjects.

The intraclass correlation was used to explain the proportion of variance due to the clustering of educators within schools. For the pre-test, 14.4% of variance was explained by this clustering and at the post-test this was 11.0%, indicating that post-test scores were less clustered at the school level. Thus, the individual level explained more of the variation in scores on the post-test.

For all models, the mean and variance of the latent scale of the pretest was fixed to identify the scale. Since the model identification was not directly implied on the intercept parameter, the estimated intercept for the pre-test was 0.002 (*SD* 0.054), and almost equal to zero. At the post-test the estimated intercept was 1.362 (*SD* 0.111). The estimated intercept at the post-test implies that, without controlling for background characteristics, data literacy ability increased compared to the pre-test. On average across educators and schools, the population score on the post-test is significantly higher. For illustrative purposes, we plotted the pre-test and post-test scores for four different schools in Fig. 3. The diamonds represent educator scores, the dotted line represents the pre-post regression line over the entire population, and the dashed line represents the pre-post regression line for the school where the fitted area is plotted as a straight, bold line. Note that the pre-post regression lines follow from the multivariate modeling approach.

The pre-post regression slopes differ over schools. For schools with positive (negative), steep pre-post slopes, the high-scoring (low-scoring) educators on the pretest showed most improvement. So, for schools with different pretest scores, a substantially different increase in the score can be expected across schools.

4.2. Adding covariates

In subsequent models, covariates were added step-wise, and variables with non-significant effects at both the time of the pre-test and post-test were excluded from the next model. Individual-level covariates were introduced before adding school-level covariates. First, educators' educational level was added, in the next model, function was included as well. In model 3 and 4, educators' gender and age were added subsequently. All individual-level covariates showed significant effects. In model 5, we controlled for cohort, which appeared to be a significant covariate. Model 6 also included the student monitoring system, but this was not significant and therefore not included in model 7, in which the proportion of master-level educators at the school level was added. This was also not significant, therefore model 5 was chosen as the final model.

When comparing model 5 with the empty model, the proportional reduction in variance at the educator level was 0.262 for the pre-test ($R^2_{pre,1}$), and 0.163 for the post-test ($R^2_{post,1}$). At level 2 (school level), the proportional reduction in variance was 0.365 for the pre-test ($R^2_{pre,2}$) and 0.248 for the post-test ($R^2_{post,2}$). This indicates that the covariates explained more variance for the pre-test than the post-test.

In this final model (see Table 7), the intercept at the pre-test was −0.077 (*SD* 0.104) and the intercept at the post-test was 1.429 (*SD* 0.185), indicating a significant increase in data literacy. The intercept can be interpreted as the data literacy ability scores for male teachers, aged 41–50, with higher education, working in a school from cohort 1. Covariates can be used to explain individual differences. We will describe the effects for each covariate separately.

Education. Educators with lower vocational education scored

significantly lower than educators with higher education, both on the pre-test as well as on the post-test. On the pre-test, people with a master's degree scored significantly higher than educators with higher education. Although the change in the effect of attaining a master's degree was not significant, the difference between educators with a higher education and master's degree was no longer significant at the time of the post-test.

Function. Academic coaches and school leaders outperformed teachers on the pre-test, whose scores were statistically comparable to the scores of people with 'other' functions at the school. On the post-test, academic coaches again scored significantly higher than teachers. The scores of the school leaders did not deviate from teachers' scores significantly, and this school leader effect itself was significantly smaller than at the time of the pre-test.

Gender. Female educators scored lower than their male colleagues, this difference was significant on the pre-test, but not on the post-test. However, the change in the effect of gender was not significant.

Age. On the pre-test, only the group of educators who are 51 years or older, scored significantly lower than the reference group of 41–50 years old. This effect remained for the post-test and was even larger during that measurement (−0.47 as opposed to −0.24 for the pre-test), but on the post-test, educators aged 30 years and younger also outperformed the age groups of 41–50 and ≥51 years old. The change in effects between the two measurement occasions was not significant for any age group.

Cohort. Educators in the second cohort scored significantly higher than educators in the first cohort. This effect was similar for both the pre-test and post-test.

Conditional on the effects of the explanatory variables, the educator's and school's pretest-posttest scores are correlated with a correlation of 0.446 and 0.474, respectively. So, the correlation between individual pretest-posttest scores was almost equal to the correlation between school-average pretest-posttest scores. The computed (conditional) pre-post regression effects (from the regression of posttest scores on pretest scores given explanatory variables) show that within schools the relationship between pretest-posttest scores was also similar to the relationship across schools. When increasing the performance on the pretest, the expected improvement on the posttest is almost identical for both educators and schools.

5. Discussion and conclusion

Data literacy is one of the preconditions for the successful implementation of data-based decision making. The [DBDM-intervention] was explicitly aimed at fulfilling preconditions in order to enable DBDM in the participating schools, and the current study was focused on investigating the effects of the intervention on the data literacy of participants.

5.1. Limitations

The main limitation of the current study is our narrow operationalization in order to measure participant's data literacy. As Gummer and Mandinach (2015) rightfully notice, when applying a narrow definition, there is a risk that the instrument is artificial and not representative of the practice any more. Although data literacy comprises many more than knowledge about the student monitoring system, interpreting the SMS output, and relating scores to benchmarks, the definition as used in this study fits the basic context of data literacy in primary education in the Netherlands. Furthermore, this definition enabled us to measure data literacy among a large group of respondents.

5.2. Conclusions

In the current study, we investigated the changes in educators' data literacy during an intensive DBDM-intervention. The first hypothesis (*participant's data literacy will improve during the intervention*) was clearly supported, results indicated a large increase in overall participants' data literacy.

By means of multivariate multi-level latent pre-post analysis, covariate effects for pre-test and post-test could be determined separately. This enabled the comparison of the covariate effects across measurement occasions and allowed us to investigate hypotheses related to 'closing the gap' based on educational level and function. Two hypotheses regarding initial data literacy were formulated. Both hypothesis 2 (*data literacy of educators with a master's degree initially will be higher*) and 3 (*initial data literacy will be higher for school leaders and academic coaches than for teachers*) were supported, this was in line with previous research (Staman, Visscher & Luyten, 2014).

At the school level, the proportion of team members that had attained a master-level education was expected to be positively associated with data literacy (*hypothesis 4)*. The effect of the proportion of master-level educators was positive, but far from significant. Apparently, having a master's degree was only beneficial to those educators themselves. There were no significant differences between data literacy in schools with small and large proportions of educators with a master's degree. Hypothesis 4 therefore was rejected.

The final two hypotheses concerned an expected decrease in the effects of educational level and function (*hypothesis 5 and 6*). Although the changes in effects of educational level were not significant, at the post-test, there was no significant effect of having a master's degree as opposed to having followed higher education. The 'data literacy gap' between attaining a master's degree and completing higher education therefore seems 'closed'. Regarding function, school leader data literacy at the post-test was comparable to teacher data literacy, whereas school leaders significantly outperformed teachers on the pre-test. The change in school leader effect was significant as well. The difference between teachers and academic coaches remained significant, in favor of the academic coaches.

Furthermore, differences in changes in data literacy across educators and student monitoring systems were explored. Overall, female educators showed lower data literacy scores than their male colleagues, and educators aged 51 years and older scored significantly lower than younger educators. At the post-test, educators aged 30 years and younger, outperformed their colleagues in the age groups 41–50 and 51 and older. Because three different SMS's are commonly used, it was investigated whether scores across systems differed. This was not the case; no significant differences were found across student monitoring systems.

This study showed that providing an intensive, long-term professional development trajectory, aimed at directly applying new skills in the context of participants' own school, can lead to significant improvements in educators' data literacy with regard to SMS data. Initial differences in scores, based on education and function, were not apparent on the post-test. Educators' data literacy was more alike, indicating that this ability can be developed and that all educators can reach the same level.

5.3. Implications

The relevance of data literacy is widely acknowledged. In the present study, the data literacy test was used as an instrument for research purposes. In future research, it would be interesting and relevant to use the test items from this study to conduct standard-

setting procedures, and to, for example, determine what would be the required starting level of data literacy for beginning teachers. The data literacy test as developed here can then be used in pre-service teacher training, to adapt their own education based on the outcomes of the tests.

Furthermore, since the ultimate aim of DBDM is to maximize student achievement, in a future study we have planned to investigate the relationship between educator data literacy and improved student achievement gains during the DBDM intervention.

Appendix A. Supplementary data

Supplementary data related to this article can be found at http://dx.doi.org/10.1016/j.tate.2017.02.015.

References

Allison, P. D. (1990). Change scores as dependent variables in regression analysis. *Sociological Methodology, 20*, 93–114.

Azevedo, C. L. N., Andrade, D. F., & Fox, J. P. (2012). A Bayesian generalized multiple group IRT model with model-fit assessment tools. *Computational Statistics and Data Analysis, 56*(12), 4399–4412. http://dx.doi.org/10.1016/j.csda.2012.03.017.

Bernhardt, V. L. (2005). Data tools for school improvement. *Educational Leadership, 62*(5), 66–69.

Bock, R. D., & Zimowski, M. F. (1997). Multiple group IRT. In W. J. van der Linden, & R. K. Hambleton (Eds.), *Handbook of modern item response theory* (pp. 433–448). New York, NY: Springer.

Bron, R., van Geel, M., & Visscher, A. (2013). *Opbrengstgericht werken op de Pabo (Data-Based Decision Making in Teacher Training)*. Report retrieved from: http://doc.utwente.nl/88199/.

Daniel, L. G., & King, D. A. (1998). Knowledge and use of testing and measurement literacy of elementary and secondary teachers. *The Journal of Educational Research, 91*(6), 331–344.

Earl, L., & Fullan, M. (2003). Using data in leadership for learning. *Cambridge Journal of Education, 33*(3), 383–394. http://doi.org/10.1080/0305764032000122023.

Earl, L., & Katz, S. (2006). *Leading schools in a data rich world*. Thousand Oaks, CA: Sage.

Fox, J. P. (2010). *Bayesian item response modeling*. New York, NY: Springer.

Fox, J. P., & Glas, C. A. W. (2001). Bayesian estimation of a multilevel IRT model using Gibbs sampling. *Psychometrika, 66*(2), 271–288.

Gotch, C. M., & French, B. F. (2013). Elementary teachers' knowledge and self-efficacy for measurement concepts. *The Teacher Educator, 48*(1), 46–57. http://doi.org/10.1080/08878730.2012.740150.

van Geel, M., Keuning, T., Visscher, A. J., & Fox, J. P. (2016). Assessing the effects of a school-wide data-based decision-making intervention onstudent achievement growth in primary schools. *American Educational Research Journal, 53*(2), 360–394.

Gummer, E., & Mandinach, E. B. (2015). Building a conceptual framework for data literacy. *Teachers College Record, 117*(4), 1–22.

Hamilton, L., Halverson, R., Jackson, S. S., Mandinach, E. B., Supovitz, J. A., & Wayman, J. C. (2009). *Using student achievement data to support instructional decision making*. Washington, DC. Retrieved from http://ies.ed.gov/ncee/wwc/pdf/practice_guides/dddm_pg_092909.pdf.

Huguet, A., Marsh, J. A., & Farrell, C. C. (2014). Building teachers' data-use capacity: Insights from strong and developing coaches. *Education Policy Analysis Archives, 22*(52), 1–31. Retrieved from epaa.asu.edu/ojs/index.php/epaa/article/download/1600/1315.

Ikemoto, G. S., & Marsh, J. A. (2007). Cutting through the "data-driven" mantra: Different conceptions of data-driven decision making. In *Evidence and decision making: Yearbook of the national society of education* (pp. 105–131). http://doi.org/10.1111/j.1744-7984.2007.00099.x.

Kamphuis, F., & Moelands, F. (2000). A student monitoring system. *Educational Measurement: Issues and Practice, 19*(4), 28–30.

Koh, K. H. (2011). Improving teachers' assessment literacy through professional development. *Teaching Education, 22*(February 2015), 255–276. http://doi.org/10.1080/10476210.2011.593164.

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., Honey, M., & Light, D. (2006). A theoretical framework for data-driven decision making. In: *Paper presented at the annual meeting of AERA* (pp. 1–18), San Francisco.

Mandinach, E. B., & Gummer, E. S. (2013a). A systemic view of implementing data literacy in educator preparation. *Educational Researcher, 42*(1), 30–37. http://doi.org/10.3102/0013189X12459803.

Mandinach, E. B., & Gummer, E. S. (2013b). Defining data literacy: A report on a convening of experts. *The Journal of Educational Research & Policy Studies, 13*(2), 6–28.

Mandinach, E. B., & Gummer, E. S. (2016). *Data Literacy for Educators: Making it count in teacher preparation and practice*. New York, NY: Teachers College Press.

Mandinach, E. B., Gummer, E. S., & Muller, R. D. (2011). *The complexities of integrating data-driven decision making into professional preparation in schools of education: It ' s harder than you think. Report from an invitational meeting* (Alexandria, VA, Portland, OR, and Washington, DC).

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, 85*, 226–245. http://doi.org/10.1080/01619561003688688.

Popham, W. (2011). Assessment literacy overlooked: A teacher educator's confession. *The Teacher Educator, 46*(April 2014), 265–273. http://doi.org/10.1080/08878730.2011.605048.

Reise, S. P., Widaman, K. F., & Pugh, R. H. (1993). Confirmatory factor analysis and item response theory: Two approaches for exploring measurement invariance. *Psychological Bulletin, 114*, 552–566. http://doi.org/10.1037/0033-2909.114.3.552.

Sklar, J. C., & Zwick, R. (2009). Multimedia presentations in educational measurement and statistics: Design considerations and instructional approaches. *Journal of Statistics Education, 17*(1991), 1–14. Retrieved from http://www.amstat.org/publications/jse/v17n3/sklar.pdf.

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

Staman, L., Visscher, A. J., & Luyten, H. (2014). The effects of professional development on the attitudes, knowledge and skills for data-driven decision making. *Studies in Educational Evaluation, 42*, 79–90. http://dx.doi.org/10.1016/j.stueduc.2013.11.002.

Stone, C. A., & Zhu, X. (2015). *Bayesian analysis of item response theory models using SAS*. Cary, NC: SAS Publications.

Supovitz, J. A. (2012). Getting at student understanding - the key to teachers' use of test data. *Teachers College Record, 114*(11), 1–29.

Verhaeghe, G., Schildkamp, K., Luyten, H., & Valcke, M. (2015). Diversity in school performance feedback systems. *School Effectiveness and School Improvement, 3453*(November), 1–27. http://doi.org/10.1080/09243453.2015.1017506.

Zwick, R., Sklar, J. C., Wakefield, G., Hamilton, C., Norman, A., & Folsom, D. (2008). Instructional Tools in Educational Measurement and Statistics (ITEMS) for school personnel: Evaluation of three web-based training modules. *Educational Measurement: Issues and Practice, 27*, 14–27. http://doi.org/10.1111/j.1745-3992.2008.00119.x.