A best-evidence meta-analysis of the effects of digital monitoring tools for teachers on student achievement

Janke M. Faber, Remco Feskens & Adrie J. Visscher

To cite this article: Janke M. Faber, Remco Feskens & Adrie J. Visscher (2023) A best-evidence meta-analysis of the effects of digital monitoring tools for teachers on student achievement, School Effectiveness and School Improvement, 34:2, 169-188, DOI: 10.1080/09243453.2022.2142247

To link to this article: https://doi.org/10.1080/09243453.2022.2142247
A best-evidence meta-analysis of the effects of digital monitoring tools for teachers on student achievement

Janke M. Faber a*, Remco Feskens b,c and Adrie J. Visscher a

aELAN Teacher Development, Faculty of Behavioural, Management and Social Sciences, University of Twente, Enschede, the Netherlands; bCito, Arnhem, the Netherlands; cCognition, Data and Education, Faculty of Behavioural, Management and Social Sciences, University of Twente, Enschede, the Netherlands

ABSTRACT

In this study the effects of the use of digital student monitoring tools for teachers (DMTs) on student achievement (primary and secondary schools, mathematics, reading, and language) were investigated through a meta-analysis (n = 14). The studies were also coded for feedback and intervention features, which resulted in three groups of combinations of DMTs and interventions. The meta-analytic findings indicate that the use of a DMT overall has a moderate effect (ES = .12) on student achievement for studies in which student achievement is measured by means of researcher-independent tests. Positive effects were also found for the use of DMTs in primary education (ES = .14), reading (ES = .17), mathematics (ES = .10), and for two groups of DMT-intervention combinations (ES = .25 and .13). Our results are encouraging but should be interpreted with caution, given the small number of studies that met our stringent inclusion criteria.

ARTICLE HISTORY

Received 24 July 2020
Accepted 27 October 2022

KEYWORDS

Meta-analysis; digital monitoring; feedback; effectiveness; teachers

Introduction and research questions

More and more, teachers and schools are using some type of digital system to store and analyze student assessment data (Faber et al., 2018; van Geel et al., 2017; Vanhoof et al., 2011; Wayman et al., 2004). Such systems typically include a set of tests to monitor students’ progress, and the software to store, organize, process, and analyze student assessment data (van Geel et al., 2017). Digital systems often provide graphics that visualize student progress, offer benchmarks to compare student achievement with (national) standards, and provide tools for aggregating (e.g., transforming student-level data into class-level data) or disaggregating (e.g., transforming class-level into student-level data) the data. The use of such tools can support teachers in the process of making grounded instructional decisions based on student progress data (Staman et al., 2017; Verhaeghe et al., 2010; Visscher & Coe, 2003), by making student assessment data more accessible than they would be otherwise.

In the literature, the presumed informative function of student assessment data for making instructional decisions and the assumed positive influence of feeding back
student achievement results to teachers are underlined (Black & Wiliam, 1998; Hellrung & Hartig, 2013). Evidence of student progress obtained by teachers on the basis of student assessment data can be considered to be feedback to teachers about the extent to which their instruction matched students’ needs (Hattie & Timperley, 2007).

The term digital monitoring tool (DMT) used in this study refers to digital tools for obtaining, storing, organizing, processing, and analyzing student assessment data that support teachers during instructional planning, by providing them with feedback based on student assessment results.

The aim of this study was to investigate whether the results of high-quality empirical studies confirm that DMTs are effective for the improvement of student achievement. Since DMTs differ, as do the interventions for implementing them, we also examined which combinations of DMT features and implementation intervention features are most effective. A best-evidence meta-analysis was conducted to summarize the results of experimental studies with teachers using a DMT in primary and secondary education. This study addressed the following questions:

- What is the effect of DMTs on student achievement?
- What DMT features and DMT implementation intervention features influence the effect of DMTs on student achievement?

**DMT features and implementation intervention features**

It was our expectation that the specific features of DMTs and of the interventions used to implement them influence the effect a DMT has on student achievement. Therefore, DMT features and implementation intervention features that are likely to matter are discussed here, starting with the features of DMTs. First of all, the frequency with which a DMT feeds back student assessment data to teachers is expected to influence the effectiveness of a DMT. Teachers seem to make a better and more rapid connection between assessment feedback and instruction if the assessment feedback is received more frequently as teachers then can relate students’ results better to how and what they taught the students, and they can also correct undesired results more quickly (Hellrung & Hartig, 2013; Koedinger et al., 2010).

The content of the feedback is also likely to be important. Hattie and Timperley (2007) argue that feedback is effective if it helps to answer three questions. First, the feedback should clarify to teachers what goals they are working on (i.e., feed up; Kluger & DeNisi, 1996; Locke & Latham, 2002). Second, feedback should provide teachers with information about progress towards these goals (feedback), and, third, feedback ideally provides teachers with information about the activities that may help making better progress towards the goals set (feed forward). Therefore, from a feedback content point of view, DMTs ideally provide information about student achievement scores to aim at (goals), information about students’ growth between assessment moments, and information about which instructional strategies might help in improving student progress. Teachers are probably better supported by concrete and detailed feedback (e.g., information about specific subject-matter topics such as division or subtraction in mathematics) than by a general score reflecting a whole subject area (Verhaeghe et al., 2010).
In addition to feedback frequency and feedback content, the level of aggregation of the feedback might also influence DMT effectiveness. DMTs often provide student-level feedback and/or class-level feedback. Class-level feedback is aggregated information on the achievement or progress of all students in a particular class, whereas student-level feedback gives information only about a single student. DMTs should include both student-level and class-level feedback, as class-level feedback is less affected by individual student characteristics than student-level feedback and is therefore more informative about the general quality of instruction.

DMTs often provide predictive feedback. In that case, the tool predicts future student and/or class performance on a national standardized assessment based on current performance. Teachers and schools can use predictive feedback as a benchmark or a goal, and goals in general have a directive function and help improve performance (Locke & Latham, 2002).

The second group of factors expected to influence the effectiveness of DMTs pertains to the interventions supporting the implementation of a DMT. Teachers will probably use a DMT better and more often if they are equipped for and supported in using the DMT. In this study, three intervention features were investigated.

The first feature is the duration of an intervention. A study by Desimone (2009) reported that although research on the length of effective teacher professionalization interventions did not indicate an exact “tipping point”, it did indicate that professional development interventions that are spread over a semester and that include a minimum of 20 hr of contact time are more likely to be effective (cf. Kennedy, 2014; Sims & Fletcher-Wood, 2021).

Second, the content of the intervention most likely will also influence how teachers use a DMT. Interventions ideally include information about and training in the technical aspects of using a DMT (Wayman et al., 2004). Teachers must learn how they can store, organize, process, and summarize student assessment data (Mandinach, 2012), especially if these steps are not automatically performed by the DMT itself. In addition to teachers learning to “push the right buttons”, teachers should also be professionalized in finding the graphs, tables, or growth models representing relevant feedback on students’ progress and instructional quality. It is likely that an intervention will be more effective if teachers also learn to translate feedback into instructional strategies (Hellrung & Hartig, 2013; Nabors Oláh et al., 2010; Wayman et al., 2011). An intervention ideally professionalizes teachers with respect to the skills and knowledge to interpret and use data, that is, with the knowledge and skills for the transformation of numbers, statistics, and other forms of output into instructional strategies that meet students’ needs (Mandinach, 2012). Furthermore, an intervention is expected to be more effective if it is adjusted to the specific context and needs of that school (Blanc et al., 2010; Verhaeghe et al., 2010).

The third group of factors influencing the effectiveness of an intervention refers to the intervention’s target group. Interventions can be targeted at one or more teachers only, at teachers and the school administration, or even at teachers, the school administration, and the school board. More collective participation in an intervention is considered to be important for effective professional development interventions, as it enables the discussion and processing of new information with colleagues, joint problem solving, and peer support (Desimone, 2009; Jansen in de Wal, 2016; Wayman et al., 2011). The use of DMTs by teachers may be better supported in schools if members of the school
administration and the school board also have the skills and knowledge to analyze and use student assessment data, and if they stimulate and support (e.g., in terms of resources) teachers in critically reviewing and using student assessment data (Blanc et al., 2010).

**Method**

In this best-evidence meta-analysis, the effects of DMTs on student achievement were estimated. We analyzed the best studies in terms of methodological quality and ecological validity and analyzed their results in a meta-analysis. The procedures followed for the literature search, the criteria for study inclusion, and the DMT analysis methods are described below.

**Literature search**

The scientific literature was searched in databases and by the consultation of international contacts. A systematic search was conducted in Educational Resources Information Center (ERIC), Web of Science, Scopus, PsycINFO, Narcis (Dutch), and International Dissertation Abstracts. In order to ensure that all relevant DMT studies would be found, research available as grey literature (not published by means of academic distribution channels) was also included in the search. This was done to prevent a potential overestimation of effects, as it is known that effective interventions are more likely to be published in peer-reviewed journals (file drawer problem).

In each database, three combinations of a wide range of keywords were used. The first combination of keywords was related to assessment: “accountability test” OR “benchmark assessment” OR “curriculum-based assessment” OR “diagnostic assessment” OR “formative assessment” OR “interim assessment” OR “standardized achievement test” OR “diagnostic test” OR “high stakes test” OR “low stakes test” OR “summative assessment”.

The second combination was related to formative assessment: “assessment for learning” OR “curriculum-based measurement” OR “data-driven” OR “data-based” OR “data analysis” OR “feedback” OR “formative evaluation” OR “assessment student progress” OR “performance-driven education” OR “progress assessment”.

The third combination of keywords was related to DMTs: “data analysis tool” OR “data reporting system” OR “electronic data management system” OR “pupil assessment system” OR “school performance feedback system” OR “student management system” OR “student assessment system” OR “student progress system”.

Each combination of keywords was also combined (using AND) with words related to the research design (“matching” OR “regression discontinuity design” OR “random” OR “experiment” OR “control group”), and with keywords related to primary and secondary education (“elementary education” OR “elementary secondary education” OR “primary education” OR “elementary school teachers” OR “elementary school” OR “Grade 1/2/3/4/5/6/7/8/9/10/11/12” OR “secondary education” OR “secondary school teachers” OR “secondary school”). Studies published before December 1990 were excluded (as finding high-quality studies prior to 1990 was considered to be very unlikely), which means that the timespan covered in the literature search was from December 1990 until May 2017.
Studies were also searched for by emailing researchers who study the use of DMTs. They were asked if they themselves knew of relevant studies, or if they knew other researchers who could be contacted for DMT studies; 126 international researchers in 24 different countries were contacted.

The title, abstract, and keywords of the articles found were screened to determine whether the effects of a DMT on student achievement had indeed been studied. The literature search and initial screening resulted in a pool of 66 articles, and the (inter)national experts pointed to another 37 potentially relevant articles. Twelve studies appeared in both groups; thus, 91 articles were screened in greater depth, based on the inclusion criteria.

**Inclusion criteria**

Two researchers independently screened all 91 studies using the following eligibility and methodological criteria:

1. Teachers in the studies had to use a digital tool that supported them during instructional planning by providing feedback based on student assessment results.
2. The interventions in the studies had to have lasted at least 12 weeks (Slavin, 2008). The intervention time was measured from the beginning of the intervention until the posttest (studies with interventions shorter than 12 weeks might be too short to detect an effect, or the observed effect may be a short-term effect that does not last). The experimental and control group in the studies had to include a total of at least 30 teachers/classes in the experimental and control groups (Kreft & de Leeuw, 2002). Maas and Hox (2004) demonstrated that multilevel studies, that is, studies which have to take into account the hierarchical structure of the data, need at least 30 observations at the group level (here the class level) in order to obtain accurate estimates of effects. To prevent that we would have to exclude too many studies, we have eased this requirement to a minimum of 20 teachers.
3. The studies had to have been conducted in a realistic school setting. The studied DMT had to be realistic in terms of its implementation in a regular school system. Large interventions outside schools or interventions that required too much time from teachers were therefore excluded.
4. Student achievement effects in the studies had to have been determined by using mathematics, reading, or language assessments to measure the dependent variable. Studies were excluded if experimenter-made assessments had been used for this purpose as effect sizes in such studies have proven to overestimate (double as large as in the case of independent tests) effect sizes (Cheung & Slavin, 2016). Experimenter-made assessments are often strongly aligned with the treatment in the experimental group, which could imply that the other subject-matter content that students have to learn in a grade is not assessed by the test. It also happens that the experimenter-made test only measures the mastery of subject-matter content to which the students in the control group never were exposed in their lessons.
5. An experimental research design had to have been used in the studies, where student achievement in the experimental group was compared with student achievement in the control group. A randomized controlled design, a propensity score matching design, or a regression discontinuity design had to have been used for the assignment of groups.
The pretest scores of the experimental and the control groups in the studies had to differ by less than 50% of a standard deviation, allowing for the comparison of the experimental and control groups. Larger pretest differences could be the result of underlying distributions/randomization problems that cannot be controlled for effectively (Shadish et al., 2002).

**Data analysis**

Effect sizes from the included studies were computed with Cohen’s $d$ (Lipsey & Wilson, 2000). Effect size $d$ was computed as the difference between the posttest mean scores of the experimental and control groups, divided by the pooled standard deviation. Regression coefficients or treatment effects and their standard errors were used if the means of the experimental and control group were not given. Study weights were used to compute the combined effect size of all included studies. Study weights were computed using the standard error of each effect size.

Mean scores belonging to different groups (e.g., separate mean scores for Grades 4 and 5) were reported in several included studies. Mean scores of different groups who used the same DMT and who had received the same intervention under different circumstances that were not relevant for this study were combined. What effects were combined is explained in Appendix 1.

The following effect sizes were estimated in this best-evidence meta-analysis: the effect of DMTs on student achievement; their effect on achievement in mathematics, reading, and language; their effect in primary education (through the sixth grade, i.e., up to and including 11–12-year-olds); and their effect in secondary education (from the seventh grade on, i.e., 12–13-year-olds and up). Furthermore, effect sizes were estimated for digital assessment tools with similar features and similar intervention features. Groups of similar assessment tools and intervention features were composed based on DMT features and intervention features.

The analyses were carried out under the assumption of a random effects model, since it was assumed that the studies had been conducted in different subpopulations. The analyses have been carried out in R (R Core Team, 2021) and more specifically the R packages metafor (Viechtbauer, 2010), esc (Lüdecke, 2019), and meta (Balduzzi et al., 2019).

**DMT and intervention implementation features coding**

DMT features and intervention features were coded by the first author based on the literature. All DMTs used in the included studies incorporated a combination of multiple DMT features and intervention features. We grouped the studies into three types of DMT and intervention features, in order to be able to investigate what DMT features and intervention combinations are most effective in improving assessment results.

The subcategories below were used for coding DMT features.

Feedback frequency subcategories were:

- once or twice a school year;
- three to seven times a school year;
- more than once a month.
Feedback content subcategories were:

- assessment scores only;
- scores and progress information;
- scores, progress information, and information on students’ mastery of specific instructional content (e.g., for spelling errors regarding the distinction between words with an \( f/v \) sound, or errors in using one or two consonants);
- scores, progress information, information on students’ mastery of specific instructional content, and instructional advice.

Feedback level was coded as:

- “yes” if student-level and class-level feedback were provided;
- “no” if only student-level feedback was provided.

Predictive feedback was coded as:

- “yes” if the DMT predicted results on national summative assessment tests;
- “no” if this was not the case.

The subcategories below were used for coding DMT implementation intervention features.

Intervention implementation frequency subcategories were:

- once or twice a year;
- three to five times a year;
- monthly.

Intervention implementation content subcategories were coded as:

- technical information (i.e., information about the technical aspects of using a DMT);
- information on how to translate DMT feedback into instruction;
- class/school support involving visiting and advising DMT users on how to use DMT feedback for improving instruction.

Intervention targets were coded as:

- only teachers;
- teachers and school principals;
- teachers, school principals, and members of the school board.

**Results**

Study selection and analysis results are described below. We will first explain what studies were not selected for the analyses, based on the inclusion criteria.
Study selection

Two raters worked independently on this and after that compared their results. In case of different results, they discussed the differences until they had reached consensus. Seventy-seven studies did not meet all inclusion criteria. The PRISMA flow diagram in Figure 1 presents an overview of the study selection process.

Most excluded studies did not meet the first criterion (i.e., teachers used a digital tool that supported them during instructional planning by providing feedback based on student assessment results, \(n = 46\)). The DMT was often used only by students, or DMTs were only a small part of an intervention. In other excluded studies no control group was used, the type of randomization was not clear, or no information was given about the comparability of the experimental and control groups (Criterion 6, \(n = 15\)). Furthermore, in some studies effects for student achievement in mathematics, reading, or language had not been measured, or experimenter-made assessments had been used (Criterion 5, \(n = 13\)). In two studies, the sample sizes did not meet the third criterion, and in one study teachers did not participate in the intervention (Criterion 4).

The remaining 14 studies were included in the best-evidence meta-analysis (Table 1). Three studies were found in the grey literature, and 11 studies were published in peer-reviewed journals. In 11 studies, students, classes, schools, or districts had been randomly assigned to experimental and control groups. In three studies, a propensity score matching design was used. Statistically significant positive effect sizes that were reported ranged from 0.06 to 0.37, and statistically nonsignificant effect sizes were reported in seven studies. A total of 493,574 students (249,739 experimental; 243,835 control) participated in the 14 studies. Five studies did not report student sample sizes due to sample

![Figure 1. PRISMA flow diagram.](image-url)
<table>
<thead>
<tr>
<th>Authors (year)</th>
<th>Title</th>
<th>Subject</th>
<th>Student age (years)</th>
<th>Study duration</th>
<th>Research design</th>
<th>n students experimental group</th>
<th>n students control group</th>
<th>Effect size</th>
</tr>
</thead>
<tbody>
<tr>
<td>1 Carlson et al. (2011)</td>
<td>A multistate district-level cluster randomized trial of the impact of data-driven reform on reading and mathematics achievement</td>
<td>Math, Reading</td>
<td>8–14</td>
<td>1 school year</td>
<td>Randomized</td>
<td>$\approx 135,439$</td>
<td>$\approx 138,074$</td>
<td>0.17 (combined)</td>
</tr>
<tr>
<td>2 Cordray et al. (2012)</td>
<td>The impact of the Measures of Academic Progress (MAP) program on student reading achievement</td>
<td>Reading</td>
<td>9–11</td>
<td>2 school years</td>
<td>Randomized</td>
<td>$1,149$ (Grade 4)</td>
<td>$765$ (Grade 4)</td>
<td>$-0.10$ (combined)</td>
</tr>
<tr>
<td>3 Faber et al. (2017)</td>
<td>The effects of a digital formative assessment tool on mathematics achievement and student motivation: Results of a randomized experiment</td>
<td>Math</td>
<td>8–9</td>
<td>5 months</td>
<td>Randomized</td>
<td>822</td>
<td>986</td>
<td>0.20</td>
</tr>
<tr>
<td>4 Faber &amp; Visscher (2018)</td>
<td>The effects of a digital formative assessment tool on spelling achievement: Results of a randomized experiment</td>
<td>Language</td>
<td>8–9</td>
<td>5 months</td>
<td>Randomized</td>
<td>619</td>
<td>986</td>
<td>0.02</td>
</tr>
<tr>
<td>5 Förster &amp; Souvignier (2014)</td>
<td>Learning progress assessment and goal setting: Effects on reading achievement, reading motivation and reading self-concept</td>
<td>Reading</td>
<td>9–10</td>
<td>6 months</td>
<td>Randomized</td>
<td>335</td>
<td>285</td>
<td>0.21</td>
</tr>
<tr>
<td>6 Konstantopoulos et al. (2013)</td>
<td>The impact of Indiana’s system of interim assessments on mathematics and reading achievement</td>
<td>Math, Reading</td>
<td>5–14</td>
<td>1 school year</td>
<td>Randomized</td>
<td>$\approx 11,344$</td>
<td>$\approx 6,587$</td>
<td>0.14 (combined)</td>
</tr>
<tr>
<td>7 Konstantopoulos et al. (2016)</td>
<td>Effects of interim assessments on student achievement: Evidence from a large-scale experiment</td>
<td>Math, Reading</td>
<td>5–14</td>
<td>1 school year</td>
<td>Randomized</td>
<td>$\approx 15,559$</td>
<td>$\approx 15,558$</td>
<td>$-0.02$ (combined)</td>
</tr>
<tr>
<td>8 May &amp; Robinson (2007)</td>
<td>A randomized evaluation of Ohio’s Personalized Assessment Reporting System (PARS)</td>
<td>Language, Math, Reading</td>
<td>15–16</td>
<td>1 school year</td>
<td>Randomized</td>
<td>$\approx 26,306$</td>
<td>$\approx 25,274$</td>
<td>$-0.92$ (combined)</td>
</tr>
<tr>
<td>9 Nunnery &amp; Ross (2007)</td>
<td>The effects of the School Renaissance program on student achievement in reading and mathematics</td>
<td>Math, Reading, Math</td>
<td>8–14</td>
<td>1 school year</td>
<td>Matching</td>
<td>416–482</td>
<td>448–510</td>
<td>0.12 (combined)</td>
</tr>
<tr>
<td>10 Ritzema (2015)</td>
<td>Professional development in data use: The effects of primary school teacher training on teaching practices and students’ mathematical proficiency</td>
<td>Math</td>
<td>7–9</td>
<td>1 school year</td>
<td>Matching</td>
<td>527</td>
<td>527</td>
<td>$-0.08$ (combined)</td>
</tr>
<tr>
<td>11 Roschelle et al. (2016)</td>
<td>Online mathematics homework increases student achievement</td>
<td>Math</td>
<td>12–13</td>
<td>2 school years</td>
<td>Randomized</td>
<td>1,621</td>
<td>1,229</td>
<td>0.18</td>
</tr>
<tr>
<td>12 Slavin et al. (2013)</td>
<td>Effects of a data-driven district reform model on state assessment outcomes</td>
<td>Math, Reading</td>
<td>8–14</td>
<td>4 school years</td>
<td>Randomized</td>
<td>$61,505$ (Grade 5)</td>
<td>$42,885$ (Grade 8)</td>
<td>0.34 (combined)</td>
</tr>
<tr>
<td>13 van der Scheer et al. (2018)</td>
<td>Effects of a data-based decision-making intervention for teachers on students’ mathematical achievement</td>
<td>Math</td>
<td>9–10</td>
<td>1 school year</td>
<td>Randomized</td>
<td>269</td>
<td>404</td>
<td>0.05</td>
</tr>
<tr>
<td>14 van Kuijk et al. (2016)</td>
<td>Goals, data use, and instruction: The effect of a teacher professional development program on reading achievement</td>
<td>Reading</td>
<td>7–9</td>
<td>1 school year</td>
<td>Matching</td>
<td>420</td>
<td>399</td>
<td>0.37</td>
</tr>
</tbody>
</table>

Note: Study duration = timespan between the pretest/start of the intervention and the posttest; sample size experimental/control group = reported in terms of student numbers if provided; if not, it was estimated (see Appendix 1).
selection and randomization of large samples at the school or the district level. The estimation of the student sample sizes for those studies is explained in Appendix 1.

In most studies, intervention effects on mathematics \((n = 10)\) and/or reading \((n = 9)\) achievement were calculated; language-related effects were investigated in two studies. The interventions in seven studies were targeted at students in primary education (through sixth grade, 11–12 years old), two studies were conducted with students in secondary education (seventh grade and up, 12–13 years old and up), and in five studies the interventions were targeted at students from both age groups.

The effects of an intervention by the Center for Data-Driven Reform in Education (CDDRE) were estimated in two studies (Studies 1 and 12 in Table 1). Effects of a DMT using standardized assessments developed by the Dutch Central Institute for Test Development (Cito) were estimated in three studies (Studies 10, 13, and 14 in Table 1). Effects of the Snappet DMT and the accompanying intervention were estimated in two studies (Studies 3 and 4 in Table 1). In these studies, the same Cito standardized assessments were used for pre- and posttests as in Studies 10, 13, and 14. The effects of mClass and Acuity and their accompanying interventions were estimated in two studies (Studies 6 and 7 in Table 1). Additionally, in three studies student achievement effects were measured for low-achieving students only, or only low-performing schools had been selected for the study (Studies 1, 11, and 12 in Table 1).

**Coding of the DMT and intervention features**

Table 2 presents an overview of the DMT and intervention implementation features in the included studies. In four studies, the feedback was given once or twice a year; in another four studies, the feedback frequency was between three to seven times a year; and in six studies, teachers received monthly or even daily feedback based on student assessments. In nine studies, the feedback included assessment scores, information on student progress between assessments, and information referring to specific subject-matter topics measured in the assessments. In two studies, the digital assessment tools provided information about assessment scores only, and in one study, the system provided feedback about student scores and student progress information. One DMT also gave instructional advice to teachers based on the analysis of the student assessment data. Most systems provided feedback at both the student and the class level \((n = 11)\). Prediction of future student performance on national standardized assessments based on actual current student performance was provided in nine DMTs.

Intervention frequency and content of the interventions showed high variability across the studies. In two studies, the intervention implementation frequency was twice a year, but teachers had their own coach whom they could consult for questions about the use of the DMT. In three studies, a train-the-trainer model was used (i.e., a few teachers at each school had two or three training sessions before the start of the school year, and these teachers trained their colleagues during the rest of the school year). In five other studies, the frequency of the intervention implementation was monthly; in most of these interventions, teachers, school principals, and/or members of the school board were trained in their schools and outside their schools in the use of their DMT. In six studies, the intervention content mostly focused on technical information about how to use the DMT; in two of those studies, teachers could consult
Table 2. The features of the digital monitoring tools (DMTs) and the interventions.

<table>
<thead>
<tr>
<th>Study</th>
<th>DMT (feedback) features</th>
<th>Intervention features</th>
<th>Group</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Frequency</td>
<td>Content</td>
<td>Class</td>
</tr>
<tr>
<td>1</td>
<td>3–7 times a year</td>
<td>scores and progress and specific contents and instruction advice</td>
<td>yes</td>
</tr>
<tr>
<td>2</td>
<td>3–7 times a year</td>
<td>and specific contents</td>
<td>no</td>
</tr>
<tr>
<td>3/4</td>
<td>more than once a month</td>
<td>and specific contents</td>
<td>yes</td>
</tr>
<tr>
<td>5</td>
<td>more than once a month</td>
<td>and specific contents</td>
<td>no</td>
</tr>
<tr>
<td>6/7</td>
<td>DMT Kindergarten–Grade 2: more than once a month</td>
<td>and progress and instruction advice</td>
<td>yes</td>
</tr>
<tr>
<td></td>
<td>DMT Grades 3–8: 3–7 times a year</td>
<td>and specific contents</td>
<td>yes</td>
</tr>
<tr>
<td>8</td>
<td>1–2 times a year</td>
<td>and specific contents</td>
<td>yes</td>
</tr>
<tr>
<td>9</td>
<td>more than once a month</td>
<td>and specific contents</td>
<td>no</td>
</tr>
<tr>
<td>10</td>
<td>1–2 times a year a</td>
<td>and specific contents</td>
<td>yes</td>
</tr>
<tr>
<td>11</td>
<td>more than once a month</td>
<td>and specific contents</td>
<td>yes</td>
</tr>
<tr>
<td>12</td>
<td>3–7 times a year</td>
<td>scores and specific contents</td>
<td>yes</td>
</tr>
<tr>
<td>13</td>
<td>1–2 times a year a</td>
<td>scores and specific contents</td>
<td>yes</td>
</tr>
<tr>
<td>14</td>
<td>1–2 times a year a</td>
<td>scores and specific contents</td>
<td>yes</td>
</tr>
</tbody>
</table>

a Teachers were also stimulated to use students’ daily work and curriculum assignments; however, the focus of the intervention seemed to be on using the results of standardized assessments as provided by the DMT.

b In each school, teachers were trained to train their colleagues. In Study 7, participants were mainly administrators, coordinators, or guidance counselors.
a coach for questions regarding instruction. In two studies, the intervention implementation included both technical information and information about how to translate data into instruction, and in six studies, the intervention implementation also included class support or school support. Goal setting and instructional differentiation (i.e., instructional material for classroom differentiation, information about the pace of instruction and instructional content, or about composing instructional groups based on varying student needs) were components of such interventions. Furthermore, some implementations included information about and training activities with respect to effective instructional strategies for mathematics or reading. Most interventions were targeted at teachers ($n = 9$); in two studies, school principals were also involved, but the focus of these interventions was also on teachers. Three interventions were targeted at district leaders, school board members, principals and key school staff, and less at teachers.

Three groups of studies with similar DMTs and intervention implementation features were composed (group membership is given in the right-hand column of Table 2). DMTs included in Group A had a low feedback frequency (3–7 times a year or 1–2 times a year), and the interventions were targeted at teachers, principals, and the school board. In two of these DMTs, the feedback content included only scores, and in one DMT the content included concrete and detailed feedback (e.g., information on specific subject-matter topics such as division or subtraction in mathematics instead of an overall score). DMTs in Group A provided class-level feedback as well as predictive feedback. The frequency of the intervention implementation activities was monthly, or two training sessions (train-the-trainer model). The implementation content included either technical information only or technical information and information about translating DMT feedback into instruction as well as class/school support.

DMTs included in Group B had a low feedback frequency (3–7 times a year or 1–2 times a year), and the interventions were targeted at teachers or at teachers and school principals. The feedback content included concrete and detailed feedback, and one DMT also provided instructional advice. Most of the systems in Group B gave class-level feedback as well as predictive feedback. The intervention implementation frequency was monthly, 3–5 times a year, or 2–3 times a year. The intervention content comprised class/school support; in one intervention, only technical information was given.

In Group C, the feedback frequency was high (more than once a month), and interventions were only targeted at teachers. The feedback content included either concrete and detailed feedback or feedback on students’ scores and students’ progress only (one DMT). Class-level feedback was provided in four of the six DMTs, and none of the DMTs provided predictive feedback. Intervention frequencies were 3–5 times a year or 1–2 times a year, and the implementation content included technical information only, translating data into instruction (one DMT), or class/school support (one DMT).

### Meta-analysis

Table 3 presents the results of the meta-analysis. As expected, the Q statistic showed significant heterogeneity among the studies, which indicates that a random model fitted the
data better than a fixed model. Figure 2 shows the distribution of effect sizes across all included studies.

As Table 3 shows, the use of a DMT overall has a positive effect on student achievement (ES = .12). Effects were found for the use of DMTs in primary (.14) and in secondary (.04; not statistically significant) education; for reading (.17), mathematics (.10), and language (.02; not statistically significant); and for interventions in Groups A (.25), B (.02; not statistically significant), and C (.13).

We conducted a sensitivity analysis of potential publication bias. Publication bias, sometimes being referred to as “the file drawer problem”, might cause bias in the estimation of the overall effect size as studies with significant (positive) effects are more likely to be published compared to studies reporting statistically nonsignificant effects. Following this reasoning, small sample size studies are more likely to be published if these report large effects, as they lack statistical power to detect small significant

---

**Table 3.** The effect sizes for the included studies based on multilevel analyses.

<table>
<thead>
<tr>
<th>Group</th>
<th>k effects</th>
<th>ES</th>
<th>LL</th>
<th>UL</th>
</tr>
</thead>
<tbody>
<tr>
<td>Overall</td>
<td>14</td>
<td>0.12*</td>
<td>0.04</td>
<td>0.19</td>
</tr>
<tr>
<td>Primary education</td>
<td>16</td>
<td>0.14*</td>
<td>0.07</td>
<td>0.22</td>
</tr>
<tr>
<td>Secondary education</td>
<td>9</td>
<td>0.04</td>
<td>−0.04</td>
<td>0.12</td>
</tr>
<tr>
<td>Mathematics</td>
<td>13</td>
<td>0.10*</td>
<td>0.02</td>
<td>0.18</td>
</tr>
<tr>
<td>Reading</td>
<td>10</td>
<td>0.17*</td>
<td>0.05</td>
<td>0.28</td>
</tr>
<tr>
<td>Language</td>
<td>2</td>
<td>0.02</td>
<td>−0.05</td>
<td>0.09</td>
</tr>
<tr>
<td>Group A</td>
<td>8</td>
<td>0.25*</td>
<td>0.12</td>
<td>0.39</td>
</tr>
<tr>
<td>Group B</td>
<td>7</td>
<td>0.02</td>
<td>−0.12</td>
<td>0.15</td>
</tr>
<tr>
<td>Group C</td>
<td>10</td>
<td>0.13*</td>
<td>0.07</td>
<td>0.19</td>
</tr>
</tbody>
</table>

Note: CI = confidence interval; k effects = the number of effects included; ES = effect size; LL = lower limit; UL = upper limit.

*p < 0.05.

---

**Figure 2.** Forest plot of the effects found in the included studies.

Note: TE = treatment effect; seTE = standard error treatment effect; SMD = standardized mean difference; CI = confidence interval.
effects. Rücker and colleagues (2011) apply this logic in a meta-analysis variant in which this type of potential bias is taken into account explicitly. The average effect size we found based on their approach is exactly the same as our average effect size. In other words, we did not find any evidence for publication bias.

Conclusion and discussion

Conclusion

The answer to the first research question (What is the effect of DMTs on student achievement?) is that the findings of our analyses indicate that the use of a DMT overall has an effect on student achievement (ES = .12). Effects for the use of DMTs in primary education were .14, and for secondary education we found an effect of .04. Effects for reading were .17, for mathematics .10, and for language .02. As far as the second research question (What DMT features and DMT implementation intervention features influence the effect of DMTs on student achievement?) is concerned, we found effects for DMTs in Group A (.25) and Group C (.13). For Group B we found a much smaller effect (.02). Each of the three groups that we made for DMT-intervention combinations was small.

Discussion

Besides the contributions this study makes to our knowledge about the impact DMTs have in realistic school settings, the study also has some limitations. First, due to our demanding inclusion criteria, the number of included studies was small. A total of 14 studies was found (10 unique DMTs) that were conducted in only three countries (8 from the USA, 5 from the Netherlands, and 1 from Germany). Given the relatively small number of studies and effect sizes on which the analyses for this study were based, caution is urged in the interpretation of these results. Since little experimental research has as yet been carried out into the effects of DMTs, it will be worthwhile to conduct further meta-analytic research into the effects of DMTs when more experimental research into this topic becomes available.

Second, the idea of a meta-analysis is to combine results of several studies that are considered combinable (Lipsey & Wilson, 2000). Even with a clearly defined definition of DMTs and the screening of studies by two researchers, the similarity of the selected DMTs may have been confounded due to our interpretations. The eye of the beholder always plays a role, especially because the descriptions in the studies are not always clear and do not always include all required information. This study limitation applies to the coding procedure, and consequently to the categorization of DMTs and interventions in Groups A, B, and C. In future research, this problem might be tackled by contacting and interviewing the authors of the studies, to verify the information obtained from the publications, and to collect the information that is missing in the publications (we do that now for another meta-analysis, and although it is time consuming, it leads to more valid information for the analyses to be conducted).

In addition to our systematic literature search, we asked experts on the topic for studies on DMTs they knew. It should be acknowledged that this is a potential source of selection bias (although our analyses do not indicate this). In future meta-analyses on DMTs, hand searching of scientific journals is therefore also advisable.
**Discussion of the findings**

Our findings for the first research question support the assumption that the use of a DMT can help teachers to improve student achievement (Staman et al., 2017; Verhaeghe et al., 2010; Visscher & Coe, 2003). To interpret the magnitude of these effects, a comparison with a benchmark is important. We preferred not to use the traditional guidelines for the classification of effect sizes as provided by Cohen (1988). We used the benchmarks developed by Kraft (2019) as these have been developed specifically for effect sizes observed in causal research into educational interventions of which the effects were measured by means of standardized instruments. Kraft studied the magnitude of the effects of 747 educational interventions in randomized studies in which intervention effects on student achievement were measured by means of standardized assessments. He found an average effect size across those studies of .16. Compared to this, the average effect found in this meta-analysis is somewhat smaller than Kraft’s average, however still moderate according to his categorization of small (smaller than .05), moderate (between .05 but smaller than .20), and large effects (.20 or larger). The effect we found for primary education is quite similar to what Kraft found for this level of education (.11). Kraft found an average effect for mathematics and for reading of .11 and .17, respectively. The effect Kraft found for reading was also found in this study. The effect Kraft found for students’ mathematic achievement is also very comparable to our meta-analysis (.10), but what we found for language is much lower (.02). In general, the groups in our meta-analysis are small: The total group of studies includes 14 studies; all other groups made were smaller. This means that our findings should be interpreted with caution.

In sum, much effort is invested in the implementation of DMTs worldwide, and the overall average effect size for high-quality studies conducted in settings that are representative for educational practice is encouraging. Our study is a first, small-scale investigation that can be used as a basis for follow-up studies to investigate why some DMTs are more successful than others, and which particular implementation characteristics make the use of DMTs more effective. In future research, it will be important to study not only the effects of DMT use but also the nature of the use of DMTs in different subject areas, in order to understand what information is especially helpful for teachers in improving their instruction, and how this is related to the nature of the subject taught.

The low number of studies in each group may have contributed to the fact that we did not find a clear pattern regarding the DMT-intervention combination features that impact effectiveness. It is remarkable that both Group A (low feedback frequency provided to teachers, principals, and school boards, class and predictive feedback, quite intensive implementation activities, focused on the provision of technical information or also more instruction-oriented implementation content) and Group C (high feedback frequency provided only to teachers, only class feedback and not very intensive intervention activity, focused mostly on technical information) had quite strong effects on student achievement. The effects for Group B (low feedback frequency, feedback for teachers and principals, concrete and detailed feedback, class and predictive feedback, varying intervention intensities and varying intervention contents across studies) was much smaller. In other words, neither the feedback frequency, the feedback target group, the feedback content, nor the intervention features revealed a clear pattern in what is decisive for impacting student achievement. It looks like it is too early to deduce such patterns.
with respect to the features of DMTs and the interventions to implement them that are
decisive for the effectiveness of DMTs in educational practice. We are in need of more
high-quality studies, and probably also of categorizations of DMTs and interventions
that are more distinctive than the ones that were used in this analysis. When more of
such studies will be available, we probably also will have more statistical power to dis-
tinguish between the groups contrasted.

Acknowledgements

This study was funded by Kennisnet, a public organization for education and ICT in the Netherlands. Kennisnet did not influence the design of the study, its results, nor the content of this article.

Disclosure statement

No potential conflict of interest was reported by the authors.

Notes on contributors

Janke M. Faber was a researcher at ELAN Teacher Development, Faculty of Behavioural, Management and Social Sciences, University of Twente, and is now a school inspector at the Dutch Inspectorate of Education.

Remco Feskens is the head of the psychometric research and services department at Cito and a guest researcher at the section Cognition, Data and Education at the Faculty Behavioural, Management and Social sciences, University of Twente. His main research interests include educational assessment and survey methodology.

Adrie J. Visscher is a full professor at the University of Twente. His research focuses on teacher professionalization and on the use and impact of feedback to teachers.

ORCID

Janke M. Faber http://orcid.org/0000-0001-8127-6831
Adrie J. Visscher http://orcid.org/0000-0001-8443-9878

References

References marked with an asterisk indicate studies included in the best-evidence meta-
analysis.


Appendix 1. How the effect sizes were calculated

Combined effect sizes were computed for the following studies (see Table 1):

- Study 2 (Cordray et al., 2012)
  Grade 4 and Grade 5 effects and standard errors were combined.

- Study 3 (Faber et al., 2017) and Study 4 (Faber & Visscher, 2018)
  The same student population was used for estimating mathematics and language effect sizes. Therefore, these effect sizes were combined for estimating the overall effect size, the primary education effect size, and the effect size of the DMTs categorized in Group C.

- Study 6 (Konstantopoulos et al., 2013) and Study 7 (Konstantopoulos et al., 2016)
  Effect sizes of the same DMTs and interventions were combined: The effects for the DMT targeting students from kindergarten through Grade 2 (K–2; mathematics and reading separately) and the effects for the DMT targeting students from Grade 3 through Grade 8 (mathematics and reading separately) were combined.

- Study 9 (Nunnery & Ross, 2007)
  Cohorts 2000–2001 and 2001–2002 effects were combined for reading for Grades 5 and 8.

For Study 1 (Carlson et al., 2011), we did not use the effect sizes mentioned in the publication but used the alternative method as suggested by the authors:

Hedges (2009) presents additional methods for calculating effect sizes in the context of a cluster randomized trial such as this one. The first approach involves calculating the ratio of the estimated treatment effect and the between-cluster variability. In math, the results of the unconditional model indicate that the standard deviation for the district-level random effect is 0.269. Dividing the point estimate of the treatment effect by this standard deviation reveals that the effect of benchmark assessments in math is equivalent to an effect size of approximately 0.21. In reading, the estimated treatment effect corresponds to an effect size of about 0.14. (Carlson et al., 2011, p. 393)

Estimated student sample sizes

Not all included studies reported student sample sizes; in those cases the student sample sizes were estimated. Below we explain how student sample sizes were estimated with the information presented in the relevant publication/article.
Study 1 (Carlson et al., 2011)
The mean number of enrolled students in the participating schools was 527 (see Table 1 in Carlson et al., 2011). The number of participating schools in the reading intervention was 524; the estimated number of enrolled students is therefore 276,148 (524*527). The number of participating schools in the mathematics intervention was 514, so that the estimated number of enrolled students is 270,878 (514*527).

Study 6 (Konstantopoulos et al., 2013)
The proportion of experimental schools was 0.63 (31/49), so the estimated student sample in the experimental condition was 11,344 (0.63*17931; see table 4 in Konstantopoulos et al., 2013). The proportion of control schools was 0.37 (18/49), so the estimated student sample in the control condition was 6,587 (0.37*17931; see table 4 in Konstantopoulos et al., 2013).

Study 7 (Konstantopoulos et al., 2016)
In this study, the effects of mClass (DMT Grades K–2) and Acuity (DFAT Grades 3–8) were analyzed separately. To estimate the sample sizes for mClass and Acuity, the reported descriptive statistics in Table 2 in Konstantopoulos et al. (2016, p. 197) were used. The total number of estimated students for mClass was 6,270 for reading and 6,249 for mathematics. The total estimated number of students for Acuity was 24,743 for reading and 24,868 for mathematics. The sample sizes for both digital systems were added for the combined effect size of mClass and Acuity.

Study 8 (May & Robinson, 2007)
A total sample size of 51,580 students is reported in table 3.1 in May and Robinson (2007). Since 51 schools participated in the experimental condition and 49 in the control condition, the estimated student sample sizes were 26,306 (51,580/100 * 51) and 25,274 (51,580/100 * 49), respectively.

Study 12 (Slavin et al., 2013)
Year 4 results were used, as the intervention had then been implemented fully. For reading and mathematics in Grade 5, a total estimated sample size of 61,505 was used. A total of 181,440 students were enrolled at the start of the intervention (59 districts), and 20 districts still participated at the end of the intervention (181,440/59 * 20). For reading and mathematics in Grade 8, a total estimated sample size of 42,885 was used. A total of 126,510 students were enrolled at the start of the intervention (59 districts), and 20 districts still participated at the end of the intervention (126,510/59 * 20).